(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 November 2022 (03.11.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/232284 A1

(51) International Patent Classification:
G10L 25/84 (2013.01) *G10L 17/06* (2013.01)
G10L 17/02 (2013.01) *G10L 17/18* (2013.01)

(21) International Application Number:
PCT/US2022/026552

(22) International Filing Date:
27 April 2022 (27.04.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/182,338 30 April 2021 (30.04.2021) US

(71) Applicant: **DOLBY LABORATORIES LICENSING CORPORATION** [US/US]; 1275 Market Street, San Francisco, California 94103 (US).

(72) Inventors: **FANELLI, Andrea**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US). **YUN, Mingqing**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US). **PANKEY, Satej Suresh**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US). **ENGEL, Nicholas Laurence**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US). **CRUM, Poppy Anne Carrie**; 5801 Scarborough Drive, Oakland, California 94611 (US).

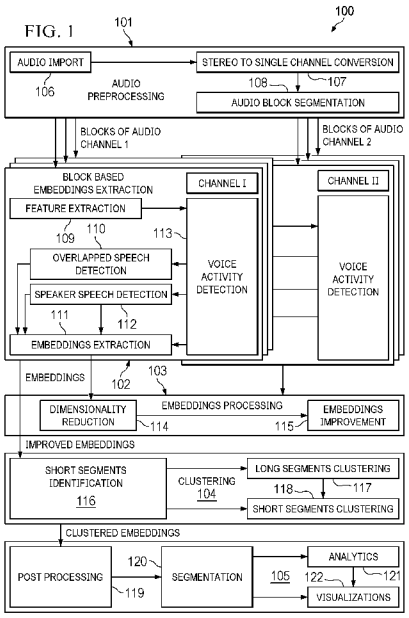
(74) Agent: **ESTES, Ernest L.** et al.; DOLBY LABORATORIES, INC., Intellectual Property Group, 1275 Market Street, San Francisco, California 94103 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) Title: SPEAKER DIARIZATION SUPPORTING EPOSODICAL CONTENT

FIG. 1



(57) Abstract: Embodiments are disclosed for speaker diarization supporting episodical content. In an embodiment, a method comprises: receiving media data including one or more utterances; dividing the media data into a plurality of blocks; identifying segments of each block of the plurality of blocks associated with a single speaker; extracting embeddings for the identified segments in accordance with a machine learning model, wherein extracting embeddings for identified segments further comprises statistically combining extracted embeddings for identified segments that correspond to a respective continuous utterance associated with a single speaker; clustering the embeddings for the identified segments into clusters; and assigning a speaker label to each of the embeddings for the identified segments in accordance with a result of the clustering. In some embodiments, a voiceprint is used to identify a speaker and the speaker identity for a speaker label.

WO 2022/232284 A1

[Continued on next page]

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

SPEAKER DIARIZATION SUPPORTING EPOSODICAL CONTENT

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority of U.S. Provisional Application 63/182,338 filed April 30, 2021.

TECHNICAL FIELD

[0002] This disclosure relates generally to audio signal processing, and more particularly to speaker diarization.

BACKGROUND

[0003] Speaker diarization is a process of partitioning an input audio stream containing speech of multiple individuals into homogeneous segments associated with each speaker. Speaker diarization is used in many applications such as understanding recorded conversations, video captioning and the like. Speaker diarization is different than speaker identification or speaker separation because speaker diarization does not require a "fingerprint" of the speaker's voice or *apriori* knowledge of the number of speakers present in the input audio stream. Additionally, speaker diarization is different than source separation because speaker diarization is not typically applied to overlapping speech.

SUMMARY

[0004] Embodiments are disclosed for speaker diarization supporting episodical content.

[0005] In some embodiments, a method comprises: receiving, with at least one processor, media data including one or more utterances; dividing, with the at least one processor, the media data into a plurality of blocks; identifying, with the at least one processor, segments of each block of the plurality of blocks associated with a single speaker; extracting, with the at least one processor, embeddings for the identified segments in accordance with a machine learning model, wherein extracting embeddings for identified segments further comprises statistically combining extracted embeddings for identified segments that

correspond to a respective continuous utterance associated with a single speaker; clustering, with the at least one processor, the embeddings for the identified segments into clusters; assigning, with the at least one processor, a speaker label to each of the embeddings for the identified segments in accordance with a result of the clustering; and outputting, with the at least one processor, speaker diarization information associated with the media data based in part on the speaker labels.

[0006] In some embodiments, before dividing the media data into a plurality of blocks, a spatial conversion on the media data. is performed.

[0007] In some embodiments, performing the spatial conversion on the media data comprises: converting a first plurality of channels of the media data into a second plurality of channels different than the first plurality of channels; and dividing the media data into a plurality blocks includes independently dividing each of the second plurality of channels into blocks.

[0008] In some embodiments, in accordance with a determination that the media data corresponds to a first media type, the machine learning model is generated from a first set of training data; and in accordance with a determination the first media data corresponds to a second media type different than the first media type, the machine learning model is generated from a second set of training data different than the first set of training data.

[0009] In some embodiments, prior to clustering, and in accordance with determining that an optimization criteria is met, the extracted embeddings for the identified segments are further optimized.

[0010] In some embodiments, prior to clustering, and in accordance with determination that an optimization criteria is not met, further optimizing the extracted embeddings for the identified segments is foregone.

[0011] In some embodiments, optimizing the extracted embeddings for identified segments includes performing at least one of dimensionality reduction of the extracted embeddings or embedding optimization of the extracted embeddings.

[0012] In some embodiments, embedding optimization includes: training the machine learning model for maximizing separability between the extracted embeddings for identified segments; and updating the extracted embeddings by applying the machine learning model for maximizing the separability between the extracted embeddings for identified segments to the extracted embeddings identified segments.

[0013] In some embodiments, the clustering comprises: for each identified segment: determining a respective length of the segment; in accordance with a determination that the

respective length of the segment is greater than a threshold length, assigning the embeddings associated with the respective identified segment according to a first clustering process; and in accordance with a determination that the respective length of the segment is not greater than a threshold length, assigning the embeddings associated with the respective identified segment according to a second clustering process different from the first clustering process.

[0014] In some embodiments, any of the foregoing methods further comprise: selecting a first clustering process from a plurality of clustering processes based in part on a determination of a quantity of distinct speakers associated with the media data.

[0015] In some embodiments, the first clustering process includes spectral clustering.

[0016] In some embodiments, the media data includes a plurality of related files.

[0017] In some embodiments, the method further comprises: selecting a plurality of the related files as the media data, wherein selecting the plurality of related files is based in part on at least one of: a content similarity associated with the plurality of related files; a metadata similarity associated with the plurality of related files; or received data corresponding to a request to process a specific set of files.

[0018] In some embodiments, the machine learning model is selected from a plurality of machine learning models in accordance with one or more properties shared by each of the plurality of related audio files.

[0019] In some embodiments, the method further comprises: computing a voiceprint distance metric between a voiceprint embedding and a centroid of each cluster; computing a distance from each centroid to each embedding belonging to that cluster; computing, for each cluster, a probability distribution of the distances of the embeddings from the centroid for that cluster; for each probability distribution, computing a probability that the voiceprint distance belongs to the probability distribution; ranking the probabilities; assigning the voiceprint to one of the clusters based on the ranking; and combining a speaker identity associated with the voiceprint with the speaker diarization information.

[0020] In some embodiments, the probability distributions are modeled as folded Gaussian distributions.

[0021] In some embodiments, the method further comprises: comparing each probability with a confidence threshold; and determining if a speaker associated with a probability has spoken based on the comparing.

[0022] In some embodiments, any of the preceding methods further comprise generating one or more analytics files or visualizations associated with the media data based in part on the assigned speaker labels.

[0023] In some embodiments, a non-transitory computer-readable storage medium stores at least one program for execution by at least one processor of an electronic device, the at least one program including instructions for performing any of the methods described above.

[0024] In some embodiments, a system comprises: at least one processor; and a memory coupled to the at least one processor storing at least one program for execution by the at least one processor, the at least one program including instructions for performing any of the methods described above.

[0025] Particular embodiments disclosed herein provide at least one or more of the following advantages: 1) an optimized architecture for speaker diarization, improved over standard diarization structures of pre-existing architectures; 2) introduction of a preprocessing step to leverage the spatial information present in stereo files before conversion to mono; 3) introduction of an embeddings optimization step to maximize embeddings separation and improve clustering based on a multi-head attention architecture or VBx clustering; 4) introduction of spectral clustering as an improved component in the pipeline; 5) introduction of a double clustering step to improve reliability of clustering and misclassification of short speaker segments; 6) ability to perform diarization on files of any length with a resulting smaller memory occupation and processing load; 7) ability to perform diarization over different files, thus allowing diarization over episodical content; 8) statistics generation, error quantifications and visualizations to easily evaluate the diarization success; and 9) a diarization pipeline that uses an input voiceprint to determine if an audio file contains speech from that person.

BRIEF DESCRIPTION OF DRAWINGS

[0026] In the drawings, specific arrangements or orderings of schematic elements, such as those representing devices, units, instruction blocks and data elements, are shown for ease of description. However, it should be understood by those skilled in the art that the specific ordering or arrangement of the schematic elements in the drawings is not meant to imply that a particular order or sequence of processing, or separation of processes, is required. Further, the inclusion of a schematic element in a drawing is not meant to imply that such element is required in all embodiments or that the features represented by such element may not be included in or combined with other elements in some embodiments.

[0027] Further, in the drawings, where connecting elements, such as solid or dashed lines or arrows, are used to illustrate a connection, relationship, or association between or among two or more other schematic elements, the absence of any such connecting elements is not meant to imply that no connection, relationship, or association can exist. In other words, some connections, relationships, or associations between elements are not shown in the drawings so as not to obscure the disclosure. In addition, for ease of illustration, a single connecting element is used to represent multiple connections, relationships or associations between elements. For example, where a connecting element represents a communication of signals, data, or instructions, it should be understood by those skilled in the art that such element represents one or multiple signal paths, as may be needed, to affect the communication.

[0028] FIG. 1 is a block diagram of a speaker diarization processing pipeline, according to an embodiment.

[0029] FIG. 2 a flow diagram illustrating how the different components of the pipeline shown in FIG. 1 work on an audio file, according to an embodiment.

[0030] FIG. 3 is an embeddings representation of a segment of audio containing four speakers, after dimensionality reduction to two dimensions, according to an embodiment.

[0031] FIG. 4A is a flow diagram of a process in the pipeline shown in FIG. 1, which provides speaker change detection and overlapped speech detection, using cross-entropy loss and binary classification models, according to an embodiment.

[0032] FIG. 4B is a flow diagram 400b in pipeline 100, which can solve the embeddings extraction process using triplet loss.

[0033] FIG. 5A is a flow diagram for processing stereo files in a diarization pipeline, according to an embodiment.

[0034] FIG. 5B is a flow diagram for an alternative processing of stereo files in a diarization pipeline, according to an embodiment.

[0035] FIG. 6 illustrates block segmentation and episodical content diarization, according to an embodiment.

[0036] FIG. 7 illustrates multi-head attention plus GE2E loss function, according to an embodiment.

[0037] FIG. 8 is a flow diagram of a double clustering process, according to an embodiment.

[0038] FIG. 9 illustrates determining distances between a voiceprint embedding and centroids of clusters identified by the diarization pipeline, according to an embodiment.

[0039] FIG. 10 illustrates approximating distances of the centroids from the embeddings belonging to the cluster as folded-gaussian distributions, according to an embodiment.

[0040] FIG. 11 illustrates ranking probability to identify which cluster has the highest probability to include the voiceprint, according to an embodiment.

[0041] FIG. 12 illustrates associating the speaker to the voiceprint talked by joining a speaker identification (ID) result with the diarization pipeline output, according to an embodiment.

[0042] FIG. 13 illustrates diarization performance for speech segments for two speakers with ground truth as a baseline, according to an embodiment.

[0043] FIG. 14 illustrates the speech segment of each speaker represented in FIG. 13, according to an embodiment.

[0044] FIG. 15 is a flow diagram of a process of speaker diarization supporting episodical content, according to an embodiment,

[0045] FIG. 16 is a block diagram of an example device architecture for implementing the features and processes described in reference to FIGS. 1-15, according to an embodiment.

[0046] The same reference symbol used in various drawings indicates like elements.

DETAILED DESCRIPTION

[0047] In the following detailed description, numerous specific details are set forth to provide a thorough understanding of the various described embodiments. It will be apparent to one of ordinary skill in the art that the various described embodiments may be practiced without these specific details. In other instances, well-known methods, procedures, components, and circuits, have not been described in detail so as not to unnecessarily obscure aspects of the embodiments. Several features are described hereafter that can each be used independently of one another or with any combination of other features.

Nomenclature

[0048] As used herein, the term "includes" and its variants are to be read as open-ended terms that mean "includes, but is not limited to." The term "or" is to be read as "and/or" unless the context clearly indicates otherwise. The term "based on" is to be read as "based at least in part on." The term "one example embodiment" and "an example embodiment" are to be read

as "at least one example embodiment." The term "another embodiment" is to be read as "at least one other embodiment." The terms "determined," "determines," or "determining" are to be read as obtaining, receiving, computing, calculating, estimating, predicting or deriving. In addition, in the following description and claims, unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skills in the art to which this disclosure belongs.

Example System

[0049] FIG. 1 is a block diagram of diarization processing pipeline 100, according to an embodiment. Diarization pipeline 100 is characterized by five main components: audio preprocessing component 101, block-based embeddings extraction component 102, embeddings processing component 103, clustering component 104 and post-processing component 105.

[0050] Diarization pipeline 100 takes as input media data (e.g., a mono or stereo audio file 106) containing utterances (e.g., speech) and returns as output corresponding segments (e.g., speech segments) with an associated tag for each speaker detected in the audio file. In an embodiment, diarization pipeline 100 also outputs analytics 121 and visualizations 120 to provide a clear way to present results of the diarization to users. For example, diarization pipeline 100 can identify the existence of multiple speakers in an audio file (e.g., 3 separate speakers in the audio file), and define the start and end time of their respective speech. In some embodiments, analytics 121 includes but is not limited to: the number of speakers in the audio file, confidence for speaker identification, time and percentage of participation to the conversation for each of the speakers, main speaker in the conversation, overlapping speech sections and conventional turns.

[0051] In some embodiments, the first component of diarization pipeline 100 is audio preprocessing component 101, which imports media data that includes speech. In some embodiments, the media data is an audio file 106. In some embodiments, audio file 106 is a mono file. In some embodiments, audio file 106 is a stereo file. In the case of a stereo file, converter 107 converts the left (L) and right (R) stereo channels of audio file 106 into two channels: channel 1 and channel 2.

[0052] In some embodiments, if the speakers are spatialized and panned over different channels then audio file 106 is converted in a manner that maximizes the spatial information that is present in channels 1 and 2. For example, if it is desired to preserve the spatial information of speaker localization, independent component analysis (ICA) can be used to generate the two channels in which spatial information is preserved and the speakers are

separated into the two channels. In some embodiments, principal component analysis (PCA) is used to perform this task. In some embodiments, deep learning or machine learning is used to perform this task. In some embodiments, audio file 106 is a stereo file that is converted to a mono file by submixing the L and R channels, or by selecting one of the two channels (e.g., discarding one of the channels, to improve processing efficiency).

[0053] After converting L and R channels into channel 1 and channel 2 to maximize the separation and retention of spatial information for panned speakers, each channel is processed by audio block segmentation component 108, which divides the audio channels 1 and 2 into blocks, which are sent to block-based embeddings extraction component 102. As used herein, a "block" is a processing unit. Each block may contain one or more speech segment(s). A "speech segment" is a time of speech where a unique speaker is talking.

[0054] Note that diarization pipeline 100 is able to process audio file 106 of any duration and can also process and perform diarization over multiple files in which the same speakers could be present. For example, in some embodiments diarization pipeline 100 can load (e.g., receive, retrieve) a stereo audio file, apply ICA to the two channels of audio file 106 to maximize and leverage the spatial information associated to panned speech, and generate channel 1 and channel 2. Diarization pipeline 100 then divides the audio file into equal length blocks (e.g., 1-second blocks), where each block is processed by downstream components of pipeline 100. In some embodiments, the blocks include blocks of multiple lengths (e.g., 0.5 seconds, 0.1 seconds, or different durations).

[0055] The blocks of each audio channel are processed by block-based embeddings extraction component 102, which performs feature extraction 109 on the blocks and applies voice activity detection (VAD) 113 to the features. In some embodiments, VAD 113 detects the speech of multiple speakers and performs overlapped speech detection 110. The speech detections are used to isolate label speech segments, i.e., portions of the blocks containing speech. The speech detections are based on a combination of results from VAD 113, speaker change detection 112 and overlapped speech detection 110.

[0056] The isolated speech segments are input into embedding extraction component 111 together with data for overlapped speech detection 110 and speaker change detection 112. Embedding extraction component 111 computes embeddings for each segment that is identified by speaker change detection 112 and the overlapped speech is discarded so that no embeddings are extracted from overlapped speech. Embedding extraction component 111 and embedding generation model to extract embeddings (e.g., a multidimensional vectorial representation of a speech segment) from the isolated speech segments. In some embodiments, embedding

extraction component 111 performs dimensionality reduction 114 on the embeddings (see FIG. 3) to generate improved embeddings 115 to be clustered by clustering component 104.

[0057] Clustering component 104 receives the improved embeddings 115 as input. Short segments identifier 116 identifies short speech segments and long speech segments which are clustered separately using long segment clustering 117 and short segment clustering 118, respectively. The embedding clusters are input into post-processing component 105. Post-processing component 105 performs post-processing 119 and segmentation 120 on the embedding clusters, which are used to generate analytics 121 and visualizations 122.

[0058] FIG. 2 illustrates in further detail how the different components of diarization pipeline 100 described above work on an audio file 106, according to an embodiment. In speech analytics, embeddings are a multidimensional vectorial representation of a segment of speech. Embeddings are generated to maximize the differentiation of different speakers in a multidimensional feature space. For example, if two speakers talk in a conversation, their generated embeddings will be multidimensional vectors that cluster in two different regions of the multidimensional space. In some embodiments, extracting embeddings includes representing single-speaker speech segments or portions of single-speaker speech segments in a multidimensional subspace. In some embodiments, extracting embeddings includes converting each single-speaker speech segment into a multidimensional vectorial representation.

[0059] The rationale behind generating embeddings is to facilitate the clustering of different speakers, thus succeeding in performing speaker diarization. Models, such as, for example, an embedding generation model are used to generate embeddings by mapping or converting speech segments to a representation in multidimensional space. The models can be trained for embeddings generation using different datasets, and by using loss functions that maximize the distance between embeddings of different speakers and minimize the distance of embeddings generated from the same speaker. In some embodiments, the embedding model is a deep neural network (DNN) trained to have a speaker-discriminative embeddings layer.

[0060] Referring to FIG. 2, in the example shown the original audio includes speech from three different speakers and ambient sound in the form of music and noise. Audio block segmentation produces a first block (Block 1) that captures speaker 1 (SPK1) speech followed by music, which is followed by speech from speaker 2 (SPK2). Block 1 is input into VAD 113, which detects the beginning and end of the SPK1 speech, resulting in a speech segment 1 (SEGM1). Multiple embeddings of speech segment 1 are generated and features are extracted from each embedding of SEGM1, and the embeddings are statistically combined (e.g., by

computing an average embedding from the multiple embeddings). An audio block is a uniform segmentation of the audio file (e.g., 1s, 2s, 3s … , etc.)

[0061] The above process is repeated for Block 2 and Block 3. For example, Block 2 generates a speech segment 4 (SEGM4), which contains speech from SPK3. Block 3 generates speech segment 7 (SEGM7), which contains speech from SPK3. Average embeddings of features are extracted from SEGM4 and SEGM7. The average embeddings for Blocks 1, 2 and 3 are then input into clustering processes (e.g., short and long segment clustering). Each cluster represents the speech from one of the 3 speakers, e.g., 3 clusters in three different regions of the multidimensional feature space.

[0062] FIG. 3 is an embeddings representation of a segment of audio containing 4 speakers, after dimensionality reduction to two dimensions using t-Distributed stochastic neighbor embedding (t-SNE), according to an embodiment. The 512-D embeddings have been reduced to 2D for visualization purposes using, for example, the t-SNE algorithm described in Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008). Each dot in FIG. 3 represents an embedding of a segment of speech. Embeddings belonging to the same speaker tend to cluster in space, and 4 distinct clusters can be identified in FIG. 3, each cluster representing a different one of the 4 speakers.

[0063] Referring again to FIG. 1, pipeline 100 includes three speech detections that are performed before the embeddings extraction: 1) VAD 113; 2) overlapped speech detection 110; and 3) speaker change detection 112. VAD 113 isolates speech segments from music or noise or other audio sources. In some embodiments, speech activity detection is performed using VAD described in the AC-4 standards. In some embodiments, VAD 113 is based on the architecture described in FIGS. 4A, 4B below. Overlapped speech detection 110 identifies regions where overlapping speech is present. Overlapping speech is normally discarded from the diarization task. Process 400a may be used to detect overlapped speech. Other architectures may also be used for overlapped speech detection, including different CNN and/or RNN architectures. Speaker change detection 112 detects a conversation turn. A conversation turn is a transition from a first speaker to a second, different speaker. Speaker change detection 112 is performed on the output of VAD 113. Process 400a may also be used for speaker change detection. Other architectures may also be used for speaker change detection, such as different CNN and/or RNN architectures.

[0064] FIG. 4A is a flow diagram of process 400a in pipeline 100, which provides speaker change detection 112 and overlapped speech detection 110, using cross-entropy loss

and binary classification models, according to some embodiments. In some embodiments, process 400a includes a convolutional neural network (CNN) architecture 401a, followed by one or more convolutional layers 402a, followed by one or more recurrent neural network (RNN) layers 403a, followed by one or more feedforward neural network (FNN) layers 404a. In some embodiments, the layers are trained against a binary classification problem (e.g., the presence or absence of speech) using suitable training techniques, such as back propagation.

[0065] FIG. 4B is a flow diagram of process 400b in diarization pipeline 100, which can solve the embeddings extraction process using triplet loss. In some embodiments, process 400b includes a convolutional neural network (CNN) layer 401b, followed by one or more convolutional layers 402b, followed by one or more recurrent layers 403b, followed by temporal pooling 404b, followed by one or more feedforward layers 404a. The output of process 400b are speech segments.

[0066] In processes 400a and 400b, CNN 401a, CNN 401b can be implemented using, for example, the SincCONV architecture described in Ravanelli, M., & Bengio, Y. (2019). Speaker Recognition from Raw Waveform with SincNet. 2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings, 1021–1028. https://doi.org/10.1109/SLT.2018.8639585. In some embodiments, the SincCONV architecture can be replaced with handcrafted features (e.g., Mel-frequency cepstral coefficients (MFCC)).

[0067] After segmenting the speech into segments where only one speaker is present (e.g., each segment does not include speech from more than one speaker as shown in FIG. 2), and isolating segments of different speakers, embeddings from speech are extracted by block-based embeddings extraction component 102. In some embodiments, x-vectors, d-vectors, i-vector or other strategies could be used for the extraction of embeddings from speech. The architecture in FIG. 4B could be used for the same purpose, after training the model on speech data to maximize the separability between different speakers. Different cost functions could be used to minimize the loss during training, such as triplet loss, cross-entropy loss, probabilistic linear discriminant analysis (PLDA) based loss, and others.

[0068] In some embodiments, embeddings are extracted over a specified window length, as shown in FIG. 2. In some embodiments, embeddings are extracted over a constant time window length (e.g., 20 ms). After the extractions, the embeddings belonging to a contiguous segment of speech from the same speaker are averaged to reduce noise and improve accuracy, as shown in FIG. 2. The number of embeddings that are averaged and the threshold that defines the maximum time window for averaging can be modified according to the length

of the input audio file 106. For long input audio files, averaging all the embeddings that are part of an entire speech segment from the same speaker is acceptable. By contrast, in the case of short input audio files, the maximum accepted time window for embeddings' averaging can be reduced, to generate an overall larger number of embeddings that will be fed to the clustering component 104 to improve clustering accuracy.

[0069] Different embeddings models can be trained for different content types using different training datasets (e.g., training data corresponding to a respective content or media type). In some embodiments, an embedding model is optimized to extract (infer) embeddings from content associated with an environment (e.g., sporting venue, concert venue, vehicle cabin, recording studio), use-case (e.g., podcast, lecture, interview, sportscast, music), a subject or topic (e.g., chemistry), an activity (e.g., commuting, walking), etc. In some embodiments, an embedding model is optimized for user generated content (e.g., amateur recordings, mobile device recordings, etc.). In some embodiments, the type of content that needs to be diarized (e.g., podcast, telephone call, educational, user generated content, etc.) is specified by an application or user of an application, and diarization pipeline 100 selects the best model for the type of content that needs to be processed in accordance the specified type of content. In some embodiments, the best model is selected according to various methodologies (e.g., a lookup table, a content analysis).

[0070] After the extraction of the embeddings 111, the embeddings' are further processed as shown in FIG. 1. In an embodiment, embeddings' processing is performed by two components: 1) embeddings dimensionality reduction component 114, and 2) embeddings optimization component 115. In some embodiments, one or both of components 114, 115 are skipped, e.g., in response to determining that the extracted embeddings are above a quality threshold.

[0071] The dimensionality reduction component 114 is implemented to improve the ability to differentiate between different speakers and to maximize the success of the clustering. The embeddings dimensionality could be reduced using PCA, t-SNE, uniform manifold approximation and projection (UMAP), PLDA or other dimensionality reduction strategies.

[0072] The embeddings optimization component 115 is described in more details in the following sections. The intent is to use a data driven approach to train a model for maximizing the separability between embeddings and further facilitate the clustering process. In some embodiments, this could be achieved using a multi-head attention plus generalized end-to-end (GE2E) loss architecture for embeddings optimization. Other architectures could also be used for this purpose.

[0073] After embeddings extraction and processing, embeddings are clustered into different speakers. The clustering component 105 performs clustering by differentiating between short speech segments 117 and long speech segments 118, as shown in FIG. 1, and as described in further detail below. Different clustering strategies can be used for short or long segment clustering. In some embodiments, an improved spectral clustering is described in, for example, Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2018).c Speaker diarization with LSTM. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April, 5239–5243.

[0074] Other clustering strategies can also be used, such as hierarchical clustering, affinity propagation, agglomerative clustering or VBx clustering. In some embodiments, the number of speakers in the audio file 106 is known or determined. For example, data can be obtained that is indicative of the number of speakers associated with the audio file is received (e.g., via an application programming interface (API) call) or derived based on analysis of the audio file using, for example, k-means clustering or other suitable clustering algorithm. In an embodiment, an elbow method can be used with the k-means clustering algorithm when the number of speakers is not known beforehand. The elbow method can include plotting a curve of an explained variation as a function of a number of clusters, and then picking the elbow of the curve as the number of clusters to use.

[0075] After clustering, post-processing 119 of the clustered embeddings is performed. Post-processing 119 is performed to remove misdetections and improve clustering accuracy. In some embodiments, a method such as a Hidden Markov Model (HMM) with predefined transitional probability between different speakers is used for this purpose. The resulting improved and postprocessed embeddings are used by segmentation component 120 to segment the audio file into segments with associated speaker label/identifiers.

[0076] Analytics 121 are generated from the clustered embeddings as output of diarization pipeline 100. Analytics 121 can include information including but not limited to: number of speakers, confidence of speaker detection, percentage and time of speaker participation in the conversation, most relevant speaker, and other analytics are generated. Diarization pipeline 100 can also generate an overview of detection accuracy in the speaker diarization task.

Example Preprocessing Block

[0077] As previously described, audio preprocessing component 101 allows diarization pipeline 100 to generate diarization output for stereo and mono input files. In case of mono

files, the audio files are treated as mono files and processed by audio block segmentation block 108, as described below. In the case of stereo files, processes 500a, 500b can be applied.

[0078] FIG. 5A is a flow diagram for processing stereo files in diarization pipeline 100, according to an embodiment. Process 500a (shown in FIG. 5A) illustrates an embodiment utilized when left channel 501a and right channel 502a are different, and speakers' voices are panned and not symmetric between right and left channels. Panned speech has spatial information that could be useful for the diarization task. For each channel 501a, 502a, spatial conversion block 503a separates the spatial information into two different channels 504a (channel 1, channel 2) to improve the success of the diarization task. In some embodiments, ICA or PCA are implemented by spatial conversion block 503a. Deep learning strategies could also be used. The remaining blocks of diarization pipeline 100 are applied separately to channel 1 and channel 2. After block processing 505a and embeddings' extraction stage 506a, embeddings are processed jointly 507a (e.g., embeddings from each channel are combined for subsequent processing), to generate a processed version of the embeddings that leverage the information associated to the spatial localization of the speakers. In some embodiments, process 500a is applied to multichannel audio (e.g., Dolby® Atmos®) instead of binaural audio.

[0079] FIG. 5B is a flow diagram for an alternative processing of stereo files in diarization pipeline 100, according to an embodiment. In some embodiments, when left and right channels 501a, 501b are identical (FIG. 4B), stereo files are converted to mono 503b and the remaining blocks of diarization pipeline 100 are applied to a single mono file as described in reference to FIG. 5A. The mono conversion is achieved by selecting one of the two channels or by downmixing.

Example Audio Block Segmentation

[0080] Audio block segmentation 108 is another step in pipeline 100 (see FIGS. 1, 2 and 6). The embeddings extracted by each speech segment (segment is a time window of speech where a unique speaker is talking), are stored and clustered after all the embeddings have been extracted from each segment over multiple audio blocks. This process introduces the possibility to process files of any length, and limits the processing and memory usage that would be required to process a file in its entirety, without affecting the performance. This process can be applied to a list of input files, thus enabling the possibility to perform diarization over episodical content.

[0081] Referring to FIG. 6, a collection of input audio files 601 from episodical content (e.g., a podcast season) are segmented by audio block segmentation 602 and fed through embedding's extraction component 603, as previously described in reference to FIG. 1. Extracted embeddings 604 are then combined together in embeddings processing 605, clustered 606 and input into postprocessed/segmentation 607. These steps allow speaker diarization to be performed over multiple audio files 106, thus enabling the possibility to compute analytics over episodical content.

Example Embeddings

[0082] Embeddings play an important role in a speaker diarization system since the embeddings are a multidimensional vectorial representation of a segment of speech. Existing solutions utilized Gaussian Mixture Modeling (GMM) based embedding extractions, i-vectors, x-vectors, d-vectors, etc. While those existing methods may be robust and show good performance on solving speaker diarization problems, they are not fully utilizing temporal information. Therefore, a new module is introduced in diarization pipeline 100 to further improve the effectiveness of embeddings by utilizing temporal information, thereby increasing the accuracy of clustering in the next step.

[0083] Several related ideas have been proposed in the literature that utilize the temporal information of speech to generate improved embeddings. For example, a long short-term memory (LSTM) based vector to sequence scoring model has been proposed, which utilizes an adjacent temporal information to generate similarity scores from embeddings. There are, however, two limitations of the LSTM structure: 1) The LSTM structure focuses more on local information and may fail in long-term dependent tasks, and 2) the LSTM structure is both time consuming and space consuming.

[0084] To address the limitations of the LSTM structure, architectures entirely based on an attention mechanism have shown promising value in sequence-to-sequence learning. Besides providing significantly faster training, attention networks demonstrate efficient modeling of long-term dependencies. For example, a positional multi-head attention model structure and triplet loss function can be used. The positional encoding is a multidimensional vector that contains information about a specific position in speech. Additionally, triplet ranking loss is utilized to learn a similarity metric using the output from multi-head attention model.

[0085] In some embodiments, a new loss function called 'End-to-End loss function' can be used to solve the speaker verification problem. GE2E training is based on processing a

large number of utterances at once, in the form of a batch that contains N speakers, and M utterances from each speaker in average. The similarity matrix $S_{ji,k}$ is defined as the scaled cosine similarities between each embedding vector e_{ji} to all centroids c_k.

$$S_{ji,k} = w \cdot \cos(e_{ji}, c_k) + b$$

[1]

[0086] Here e_{ji} represents the embedding vector of the j^{th} speaker's i^{th} utterance (1<=i<=M, j<=N), c_k represents the centroid of the embedding vectors where k <= N.

[0087] Since the speaker diarization problem is also trying to find a similarity matrix using the extracted embedding, and then perform clustering algorithm based on the similarity matrix, the constraint on M utterance in Equation [1] can be removed and a similar equation can be used to compute loss, because the speaker can talk unlimited number of times in speaker diarization problem. Similar as Equation [1], e_{ji} represents the embedding vector of the j^{th} speaker's i^{th} utterance (i>=1, j<=N), c_k represents the centroid of the embedding vectors where k <=N. Then if a softmax function is used on Equation [1], the loss on each embedding vector e_{ji} could be defined as:

$$\mathcal{L}(e_{ji}) = -S_{ji,k} + log \sum_{k=1}^{N} \exp(S_{ji,k}) .$$

[2]

[0088] FIG. 7 illustrates multi-head attention plus GE2E loss function, according to an embodiment. As shown in FIG. 7, the overall embedding optimization module 700 sends input embeddings 701 into a multi-head attention model. Positional encoding 702 is applied for each time step, which means the training process uses temporal pooling 706 to take temporal information into consideration in the training process. The training process minimizes the loss function 707 (e.g., GE2E loss function shown in Equation [2]). The loss function 707 pushes each embedding vector close to its centroid and pulls it away from all other centroids. During testing, the weights and bias trained within the multi-head attention model are utilized to predict optimized embeddings for each input and form a similarity matrix 708, which input into clustering 709. Since it considers the temporal information, it should be able to struct a better similarity matrix compares to the original embedding.

Example Clustering Techniques

[0089] Spectral clustering is a standard clustering algorithm that has been used in multiple tasks, including diarization. In diarization pipeline 100, a modified version of spectral clustering is used. FIG. 1 illustrates an embodiment combining a modified version of spectral clustering with diarization pipeline 100 described in reference to FIG. 1. Modified spectral clustering is one example clustering strategy that is supported by diarization pipeline 100. In some embodiments, other clustering strategies can be used, including but not limited to: hierarchical clustering, k-means, affinity propagation, or agglomerative clustering can be used.

[0090] As previously described, to remove noise and improve embeddings' estimation accuracy, in some embodiments the embeddings generated over a speech segment where a single speaker is present are averaged. For this reason, a short duration speech segment may not be informative enough for the characterization of a speaker. This might be due to embeddings inaccuracies that would be averaged out in a long duration speech segment. Hence, a double step clustering approach is proposed as described in reference to FIG. 8.

[0091] In some embodiments, Bayesian HMM clustering of x-vector sequences(VBx) can be used in diarization pipeline 100, which is described in Landini, Federico, Ján Profant, Mireia Diez, and Lukáš Burget. "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks." *Computer Speech & Language*71 (2022): 101254. In diarization pipeline 100, VBx is modified to take an x-vector and other embeddings.

[0092] FIG. 8 is a flow diagram of a double clustering process, according to an embodiment. As shown in FIG. 8, embeddings are extracted for each speech segment 801. The speech segments are separated into long-duration speech segments 802 and short-duration speech segments 803 based on a duration threshold. Long-duration speech segments 802 are clustered first 803, and the center of each cluster is calculated 804. The cosine distance (e.g., Eisen cosine correlation distance) between each of the short-duration speech segment and each cluster is computed 805, and the short-duration speech segments 803 are assigned to the closest cluster. Other metrics of distance could be used (e.g., Euclidean, Manhattan, Pearson, Spearman, Kendall correlation distances), or other strategies can be employed to assign each short speech segments 803 to the optimal cluster. The overall clustering result is a combination of the clustering results of short-duration speech segments 803 and long-duration speech segments 802.

Example Postprocessing

[0093] Another component in processing pipeline 100 is postprocessing component 105. Postprocessing component 105 is introduced in pipeline 100 to reduce errors in segmentation and clustering. Postprocessing 105 analyzes annotations generated by the clustering, and identifies and corrects possible errors. Several postprocessing strategies can be used in pipeline 100. For example, an HMM with an associated probability of speaker change could be used. Other algorithms that leverage the temporal relationship between annotations can be used for the same purpose.

Example Speaker Identification

[0094] Speaker identification (speakerID) is the ability to understand the speaker identity based on an input voiceprint. In some embodiments, diarization pipeline 100 uses an input voiceprint (e.g., 10-20 seconds of a person talking) to determine if an audio file contains speech from that person. Additionally, based on the knowledge of voiceprints of multiple speakers, speakerID labels can be assigned to the output of diarization pipeline 100. As described above, diarization pipeline 100 computes embeddings for each speech segment, and subsequently clusters embeddings into a limited number of clusters.

[0095] FIGS. 9 and 10 illustrate how the distances of reference points from the embeddings belonging to the cluster are approximated as folded-gaussian distributions, according to an embodiment. Reference points can be centroids or any other suitable reference point, such medoids in partition around medoids (PAM) clustering. As shown in FIG. 9, the first step of the speakerID process is computing a voiceprint distance metric, which in this example is a cosine distance metric, d_vp_i, between the voiceprint *vp* (905) and each reference point of each cluster (e.g., centroids c_i (for clusters 901-904)), where *i* is an index identifying the clusters, going from 1 to N, where N is the number of clusters. In the example shown, N=4 clusters representing four main speakers.

[0096] For each cluster, the cosine distance from the reference point of the cluster to each embedding belonging to that cluster is computed. For each cluster 901-904, the distribution of the distances of the embeddings from the reference point belonging to that cluster are also computed. In some embodiments, the distributions are modeled as folded Gaussian distributions (p_i), as shown in FIG. 10. Subsequently, the probability that the voiceprint distances belong to the corresponding distribution are computed. The probabilities can be ranked and assigned the voiceprint to the corresponding cluster, as illustrated in FIG. 11, where Spk1 in this example has the highest probability. By defining a confidence threshold, the probability that the voiceprint's person spoke at all in the audio file can be better

understood. Understanding when the speaker associated to a voiceprint spoke can be determined by joining the speakerID result with the output of diarization pipeline 100, as illustrated in FIG. 12. In the example shown in FIG. 12, the speakerID result is used to identify speech segments 1201-1 through 1201-4, where Spk1 is speaking.

Example Analytics/Visualizations

[0097] Diarization pipeline 100 generates analytics, metrics and visualizations as output. The output provides a user with an objective and clear quantification of the output of diarization pipeline 100. In some embodiments, analytics and visualizations are generated for a single file, for multiple episodical files, and/or for entire datasets. Table 1 and Table 2 below summarize example analytics that can be generated by diarization pipeline 100. In some embodiments, a visualization includes causing the display of a first object representing a first speaker, a second object representing a respective speaker, and a connector object connecting the first and second object that varies in appearance (length, size, color, shape, style, etc.) according to one or more statistics derived from the clustered embeddings.

Table 1 – Example Single File Analytics Summary

Name	Purpose
Number of talkers	Get total number of talkers in the conversation
Confidence of each talker detection	Understand how confident we about each detected talker in the conversation are
Percentage of each talker participation in the conversation	Understand who is the dominant talker / How much time does each talker talks in the conversation
Speech segments for each talker	Know start and end time of each speech segment for each talker
Most relevant talker	Understand who is the main speaker in a conversation, host of a podcast, main presenter, etc.
Longest monologue for each talker	Understand the longest monologue for each talker

Conversation changes	Identify when there is a change of speaker in conversation
Speaker overlaps	Identify when two speakers speak at the same time
Politeness Level	Quantify how polite a speaker is, based on number of interruptions of other speakers
Number of silences	Quantifies silences in the conversation.

Table 2 - Example Episodical Analytics Summary

Name	Purpose
Number of talkers	Get total number of talkers in the episodical content
Conversation time for each talker	How much time does each talker talks in the conversation, over multiple episodes.
Interaction between talkers	Define interaction of talkers in the episodes of the season using interaction graph (Figure 8).
Main talker per episode	Define the main talker for each episode
Expansion of single file analytics to full season	All the analytics described for single file, could be used to describe seasonal content as well.
...	

[0098] Additionally, the performance diarization pipeline 100 can be evaluated. With a diarization ground truth, the evaluation metrics in Table 3 can be used to generate a report that allows the user to better evaluate the performance of diarization pipeline 100.

Table 3 - Example Evaluation Metrics Summary

Module	Evaluation metrics
Speech activity detection	Detection cost function
	Detection error rate
	Detection accuracy
	Detection precision

	Detection recall
Segmentation	Segmentation purity
	Segmentation coverage
Diarization	Diarization error rate
	Jaccard error rate
	Diarization purity
	Diarization coverage
	Diarization completeness
	Diarization homogeneity
Speaker focused evaluation	Number of speakers
	Speaker duration
	Precision for each speaker
	Recall for each speaker
	Accuracy for each speaker

[0099] In some embodiments, diarization pipeline 100 includes a visualization module in to help users understand the diarization performance in a more efficient way. For example, FIG. 13 shows the percentage of time that each speaker talks, which is helpful for understanding the relative duration of each speaker and how much of the speech has been correctly detected. FIG. 14 shows the speech segment of each speaker, with ground truth shown as a baseline, which is helpful for understanding the performance of diarization pipeline 100 more intuitively.

Example Process

[00100] FIG. 15 is a flow diagram of process 1500 of context aware audio processing, according to an embodiment. Process 1500 can be implemented using, for example, device architecture 1500 described in reference to FIG. 16.

[00101] Process 1500 includes the steps of receiving media data including one or more utterances (1501), dividing the media data into a plurality of blocks (1502), identifying segments of each block of the plurality of blocks associated with a single speaker (1503), extracting embeddings for the identified segments in accordance with a machine learning

model (1504), clustering the embeddings for the identified segments into clusters (1505), and assigning a speaker label to each of the embeddings for the identified segments in accordance with a result of the clustering (1506). Each of these steps were previously described in detail above in reference to FIG. 1-14.

Example System Architecture

[00102] FIG. 16 shows a block diagram of an example system 1600 suitable for implementing example embodiments described in reference to FIGS. 1-15. System 1600 includes central processing unit (CPU) 1601 which is capable of performing various processes in accordance with a program stored in, for example, read only memory (ROM) 1602 or a program loaded from, for example, storage unit 1608 to random access memory (RAM) 1603. In the RAM 1603, the data required when CPU 1601 performs the various processes is also stored, as required. CPU 1601, ROM 1602 and RAM 1603 are connected to one another via bus 1604. Input/output (I/O) interface 1605 is also connected to bus 1604.

[00103] The following components are connected to I/O interface 1605: input unit 1606, that may include a keyboard, a mouse, or the like; output unit 1607 that may include a display such as a liquid crystal display (LCD) and one or more speakers; storage unit 1608 including a hard disk, or another suitable storage device; and communication unit 1609 including a network interface card such as a network card (e.g., wired or wireless).

[00104] In some embodiments, input unit 1606 includes one or more microphones in different positions (depending on the host device) enabling capture of audio signals in various formats (e.g., mono, stereo, spatial, immersive, and other suitable formats).

[00105] In some embodiments, output unit 1607 include systems with various number of speakers. Output unit 1607 can render audio signals in various formats (e.g., mono, stereo, immersive, binaural, and other suitable formats).

[00106] Communication unit 1609 is configured to communicate with other devices (e.g., via a network). Drive 1610 is also connected to I/O interface 1605, as required. Removable medium 1611, such as a magnetic disk, an optical disk, a magneto-optical disk, a flash drive or another suitable removable medium is mounted on drive 1610, so that a computer program read therefrom is installed into storage unit 1608, as required. A person skilled in the art would understand that although system 1600 is described as including the above-described components, in real applications, it is possible to add, remove, and/or replace some of these components and all these modifications or alteration all fall within the scope of the present disclosure.

[00107] In accordance with example embodiments of the present disclosure, the processes described above may be implemented as computer software programs or on a computer-readable storage medium. For example, embodiments of the present disclosure include a computer program product including a computer program tangibly embodied on a machine readable medium, the computer program including program code for performing methods. In such embodiments, the computer program may be downloaded and mounted from the network via the communication unit 709, and/or installed from removable medium 1611, as shown in FIG. 16.

[00108] Generally, various example embodiments of the present disclosure may be implemented in hardware or special purpose circuits (e.g., control circuitry), software, logic or any combination thereof. For example, the units discussed above can be executed by control circuitry (e.g., a CPU in combination with other components of FIG. 16), thus, the control circuitry may be performing the actions described in this disclosure. Some aspects may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device (e.g., control circuitry). While various aspects of the example embodiments of the present disclosure are illustrated and described as block diagrams, flowcharts, or using some other pictorial representation, it will be appreciated that the blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

[00109] Additionally, various blocks shown in the flowcharts may be viewed as method steps, and/or as operations that result from operation of computer program code, and/or as a plurality of coupled logic circuit elements constructed to carry out the associated function(s). For example, embodiments of the present disclosure include a computer program product including a computer program tangibly embodied on a machine readable medium, the computer program containing program codes configured to carry out the methods as described above.

[00110] In the context of the disclosure, a machine readable medium may be any tangible medium that may contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device. The machine readable medium may be a machine readable signal medium or a machine readable storage medium. A machine readable medium may be non-transitory and may include but not limited to an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples of the machine readable storage

medium would include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing.

[00111] Computer program code for carrying out methods of the present disclosure may be written in any combination of one or more programming languages. These computer program codes may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus that has control circuitry, such that the program codes, when executed by the processor of the computer or other programmable data processing apparatus, cause the functions/operations specified in the flowcharts and/or block diagrams to be implemented. The program code may execute entirely on a computer, partly on the computer, as a stand-alone software package, partly on the computer and partly on a remote computer or entirely on the remote computer or server or distributed over one or more remote computers and/or servers.

[00112] While this document contains many specific embodiment details, these should not be construed as limitations on the scope of what may be claimed, but rather as descriptions of features that may be specific to particular embodiments. Certain features that are described in this specification in the context of separate embodiments can also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment can also be implemented in multiple embodiments separately or in any suitable sub combination. Moreover, although features may be described above as acting in certain combinations and even initially claimed as such, one or more features from a claimed combination can, in some cases, be excised from the combination, and the claimed combination may be directed to a sub combination or variation of a sub combination. Logic flows depicted in the figures do not require the particular order shown, or sequential order, to achieve desirable results. In addition, other steps may be provided, or steps may be eliminated, from the described flows, and other components may be added to, or removed from, the described systems. Accordingly, other embodiments are within the scope of the following claims.

CLAIMS

1. A method comprising:

receiving, with at least one processor, media data including one or more utterances;

dividing, with the at least one processor, the media data into a plurality of blocks;

identifying, with the at least one processor, segments of each block of the plurality of blocks associated with a single speaker;

extracting, with the at least one processor, embeddings for the identified segments in accordance with a machine learning model, wherein extracting embeddings for identified segments further comprises statistically combining extracted embeddings for identified segments that correspond to a respective continuous utterance associated with a single speaker;

clustering, with the at least one processor, the embeddings for the identified segments into clusters;

assigning, with the at least one processor, a speaker label to at least one of the embeddings for the identified segments in accordance with a result of the clustering; and

outputting, with the at least one processor, speaker diarization information associated with the media data based in part on the speaker labels.

2. The method of claim 1, further comprising:

before dividing the media data into a plurality of blocks, performing, with the at least one processor, a spatial conversion on the media data.

3. The method of claim 2, wherein performing the spatial conversion on the media data comprises:

converting a first plurality of channels of the media data into a second plurality of channels different than the first plurality of channels; and

dividing the media data into a plurality blocks includes independently dividing each of the second plurality of channels into blocks.

4. The method of any of claims 1-3, further comprising:

in accordance with a determination that the media data corresponds to a first media type, the machine learning model is generated from a first set of training data; and

in accordance with a determination the first media data corresponds to a second media type different than the first media type, the machine learning model is generated from a second set of training data different than the first set of training data.

5. The method of any of claims 1-4, further comprising:

prior to clustering, and in accordance with determining that an optimization criteria is met, further optimizing the extracted embeddings for the identified segments.

6. The method of claim 5, further comprising:

prior to clustering, and in accordance with determination that an optimization criteria is not met, foregoing further optimizing the extracted embeddings for the identified segments.

7. The method of any of claims 5-6, wherein optimizing the extracted embeddings for identified segments includes performing at least one of dimensionality reduction of the extracted embeddings or embedding optimization of the extracted embeddings.

8. The method of claim 7, wherein embedding optimization includes:

training the machine learning model for maximizing separability between the extracted embeddings for identified segments; and

updating the extracted embeddings by applying the machine learning model for maximizing the separability between the extracted embeddings for identified segments to the extracted embeddings identified segments.

9. The method of any of claims 1-8, wherein the clustering comprises:

for each identified segment:

determining a respective length of the segment;

in accordance with a determination that the respective length of the segment is greater than a threshold length, assigning the embeddings associated with the respective identified segment according to a first clustering process; and

in accordance with a determination that the respective length of the segment is not greater than a threshold length, assigning the embeddings associated with the respective identified segment according to a second clustering process different from the first clustering process.

10. The method of any of claims 1-9, further comprising:

selecting a first clustering process from a plurality of clustering processes based in part on a determination of a quantity of distinct speakers associated with the media data.

11. The method of any of claims 1-10, wherein the first clustering process includes spectral clustering.

12. The method of any of claims 1-10, wherein the media data includes a plurality of related files.

13. The method of claim 12, further comprising:

selecting a plurality of the related files as the media data, wherein selecting the plurality of related files is based in part on at least one of:

a content similarity associated with the plurality of related files;

a metadata similarity associated with the plurality of related files; or

received data corresponding to a request to process a specific set of files.

14. The method of any of claims 12-13, wherein the machine learning model is selected from a plurality of machine learning models in accordance with one or more properties shared by each of the plurality of related audio files.

15. The method of claim 1, further comprising:

computing a voiceprint distance metric between a voiceprint embedding and a reference point of each cluster;

computing a distance from each reference point to each embedding belonging to that cluster;

computing, for each cluster, a probability distribution of the distances of the embeddings from the reference point for that cluster;

for each probability distribution, computing a probability that the voiceprint distance belongs to the probability distribution;.

ranking the probabilities;

assigning the voiceprint to one of the clusters based on the ranking; and

combining a speaker identity associated with the voiceprint with the speaker diarization information.

16. The method of claim 15, wherein the probability distributions are modeled as folded Gaussian distributions.

17. The method of claim 15, further comprising:

comparing each probability with a confidence threshold; and

determining if a speaker associated with a probability has spoken based on the comparing.

18. The method of any of claims 1-17, further comprising:

generating one or more analytics files or visualizations associated with the media data based in part on the assigned speaker labels.

19. A method comprising:

receiving, with at least one processor, media data including one or more utterances;

dividing, with the at least one processor, the media data into a plurality of blocks;

identifying, with the at least one processor, segments of each block of the plurality of blocks associated with a single speaker;

extracting, with the at least one processor, embeddings for the identified segments in accordance with a machine learning model;

spectral clustering, with the at least one processor, the embeddings for the identified segments into clusters;

assigning, with the at least one processor, a speaker label to at least one of the embeddings for the identified segments in accordance with a result of the clustering; and

outputting, with the at least one processor, speaker diarization information associated with the media data based in part on the speaker labels.

20. The method of claim 19, further comprising:

before dividing the media data into a plurality of blocks, performing, with the at least one processor, a spatial conversion on the media data.

21. The method of claim 20, wherein performing the spatial conversion on the media data comprises:

converting a first plurality of channels of the media data into a second plurality of channels different than the first plurality of channels; and

dividing the media data into a plurality blocks includes independently dividing each of the second plurality of channels into blocks.

22. A non-transitory computer-readable storage medium storing at least one program for execution by at least one processor of an electronic device, the at least one program including instructions for performing the methods of any of claims 1-21.

23. A system comprising:

at least one processor; and

a memory coupled to the at least one processor storing at least one program for execution by the at least one processor, the at least one program including instructions for performing the methods of any of claims 1-21.

FIG. 1

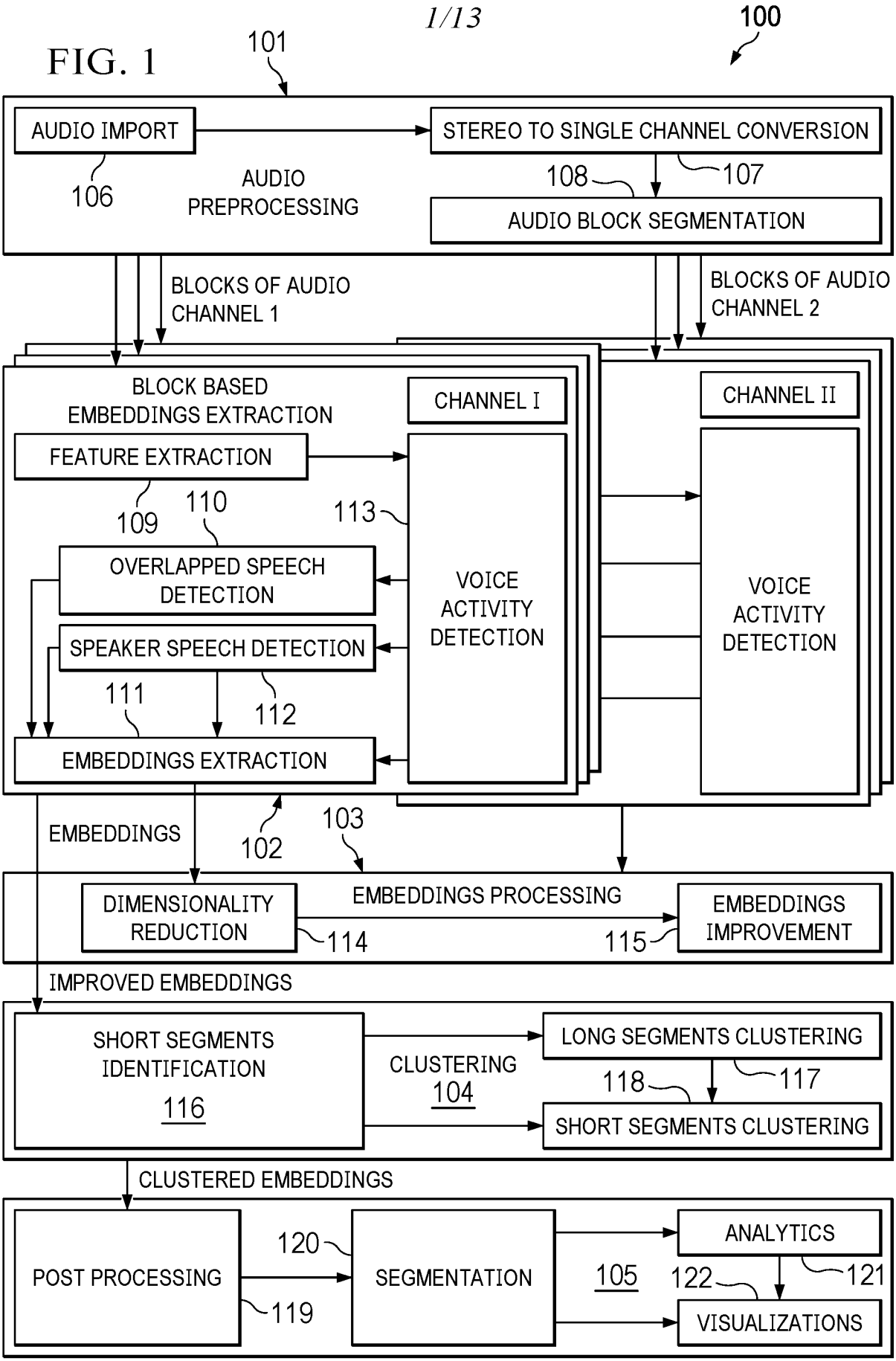
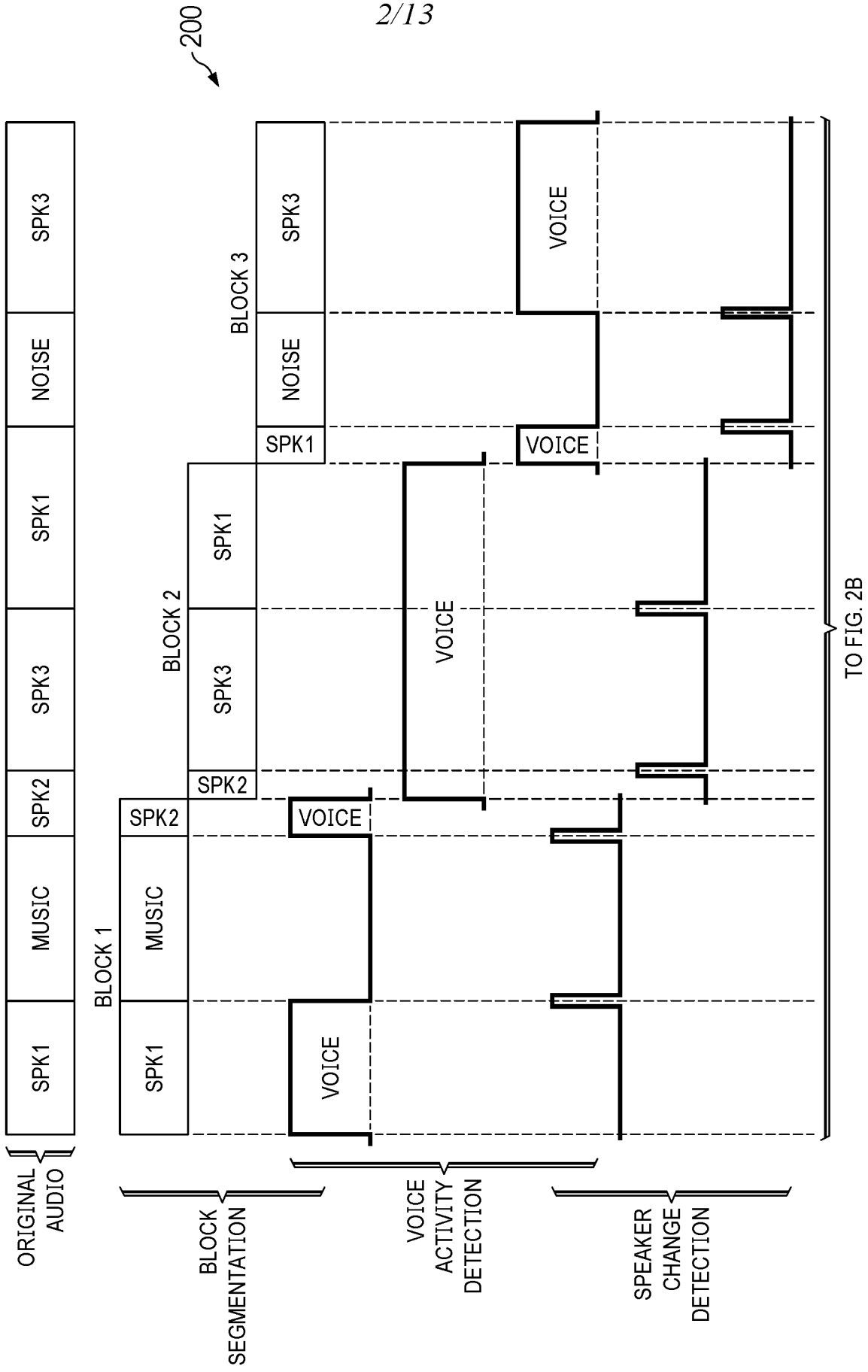

100
101
AUDIO IMPORT
106
STEREO TO SINGLE CHANNEL CONVERSION
108
107
AUDIO BLOCK SEGMENTATION
AUDIO PREPROCESSING
BLOCKS OF AUDIO CHANNEL 1
BLOCKS OF AUDIO CHANNEL 2
BLOCK BASED EMBEDDINGS EXTRACTION
CHANNEL I
CHANNEL II
FEATURE EXTRACTION
109
110
113
OVERLAPPED SPEECH DETECTION
SPEAKER SPEECH DETECTION
VOICE ACTIVITY DETECTION
VOICE ACTIVITY DETECTION
111
112
EMBEDDINGS EXTRACTION
EMBEDDINGS
102
103
DIMENSIONALITY REDUCTION
114
EMBEDDINGS PROCESSING
115
EMBEDDINGS IMPROVEMENT
IMPROVED EMBEDDINGS
SHORT SEGMENTS IDENTIFICATION
116
CLUSTERING
104
LONG SEGMENTS CLUSTERING
118
117
SHORT SEGMENTS CLUSTERING
CLUSTERED EMBEDDINGS
POST PROCESSING
120
119
SEGMENTATION
105
ANALYTICS
122
121
VISUALIZATIONS

FIG. 2A


200
ORIGINAL AUDIO
SPK1
MUSIC
SPK2
SPK3
SPK1
NOISE
SPK3
BLOCK 1
SPK1
MUSIC
SPK2
BLOCK 2
SPK2
SPK3
SPK1
BLOCK 3
SPK1
NOISE
SPK3
VOICE
VOICE
VOICE
VOICE
BLOCK SEGMENTATION
VOICE ACTIVITY DETECTION
SPEAKER CHANGE DETECTION
TO FIG. 2B

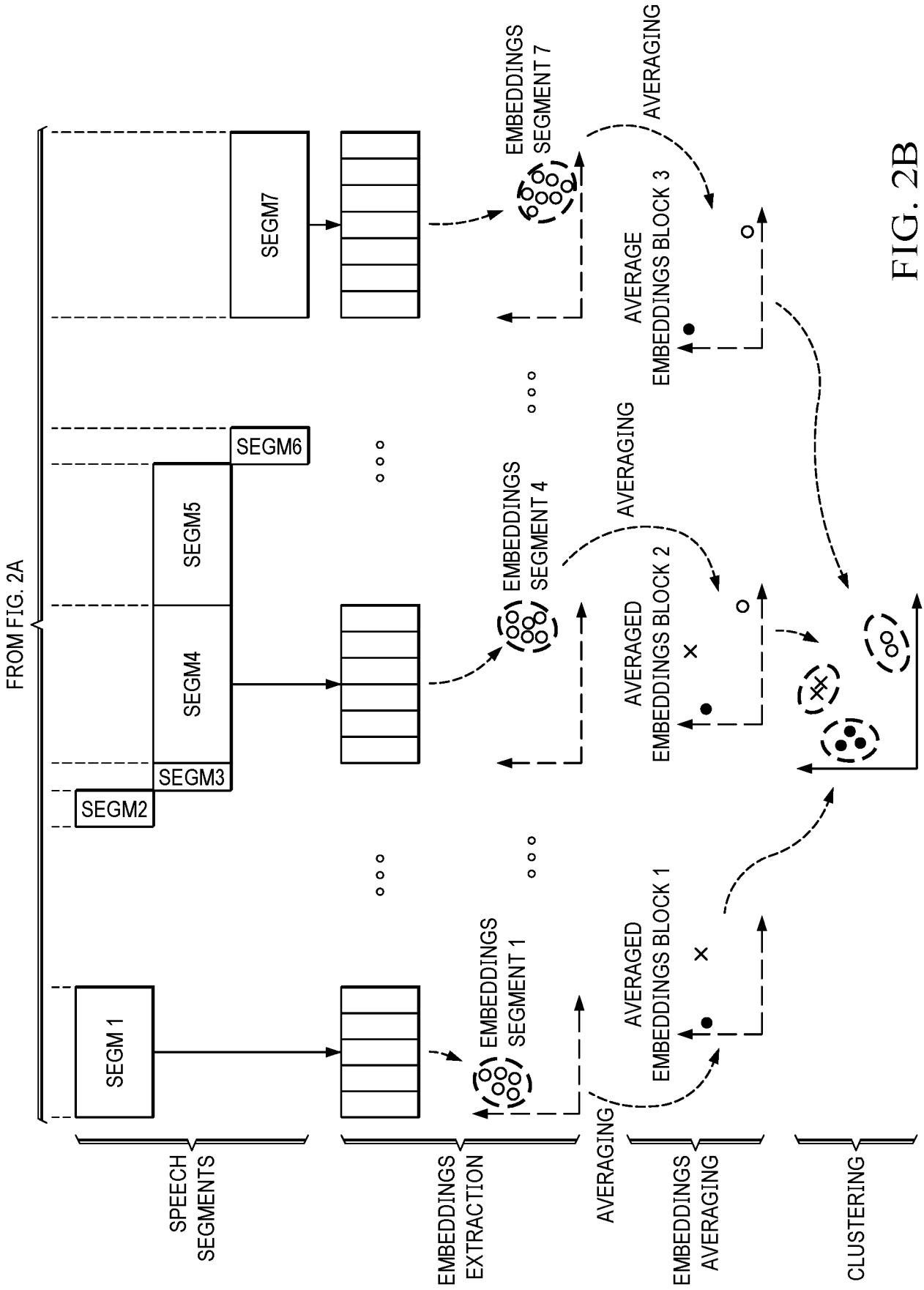

FROM FIG. 2A
SEGM 1
SEGM2
SEGM3
SEGM4
SEGM5
SEGM6
SEGM7
EMBEDDINGS SEGMENT 1
EMBEDDINGS SEGMENT 4
EMBEDDINGS SEGMENT 7
AVERAGING
AVERAGING
AVERAGED EMBEDDINGS BLOCK 1
AVERAGED EMBEDDINGS BLOCK 2
AVERAGE EMBEDDINGS BLOCK 3
SPEECH SEGMENTS
EMBEDDINGS EXTRACTION
AVERAGING
EMBEDDINGS AVERAGING
CLUSTERING

FIG. 2B

FIG. 3

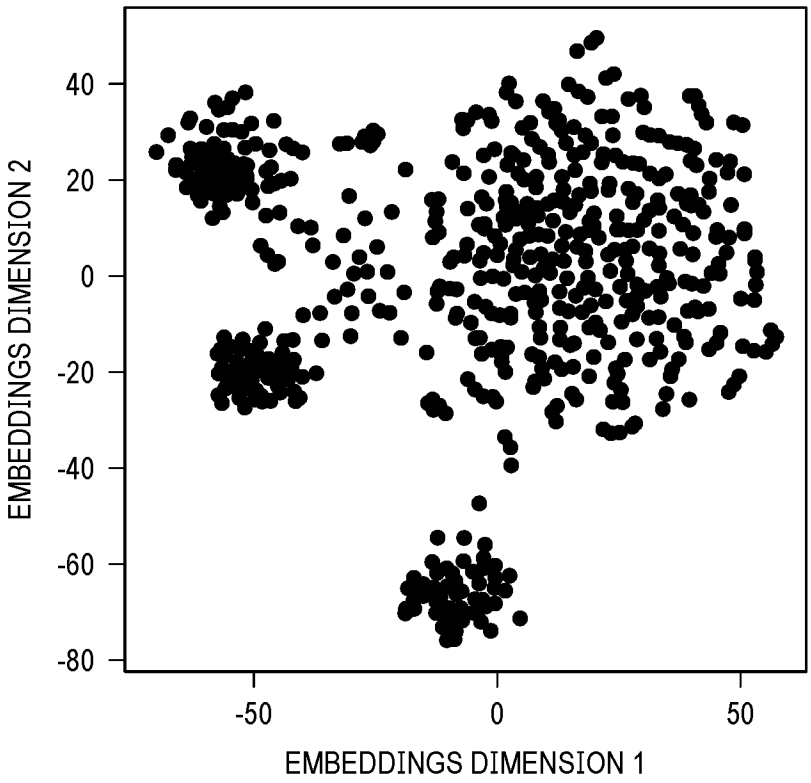


FIG. 4A

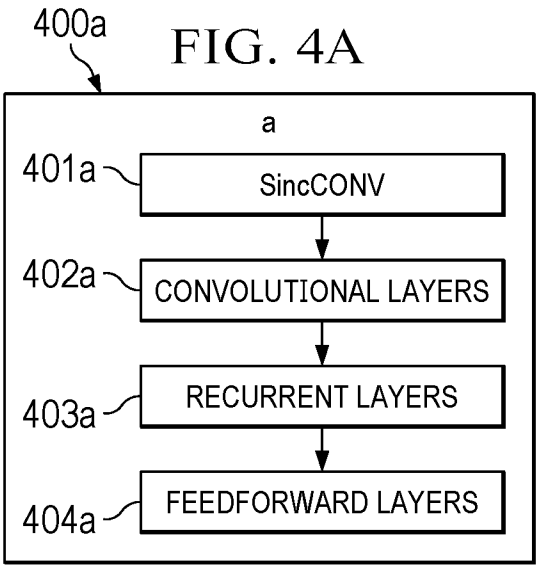
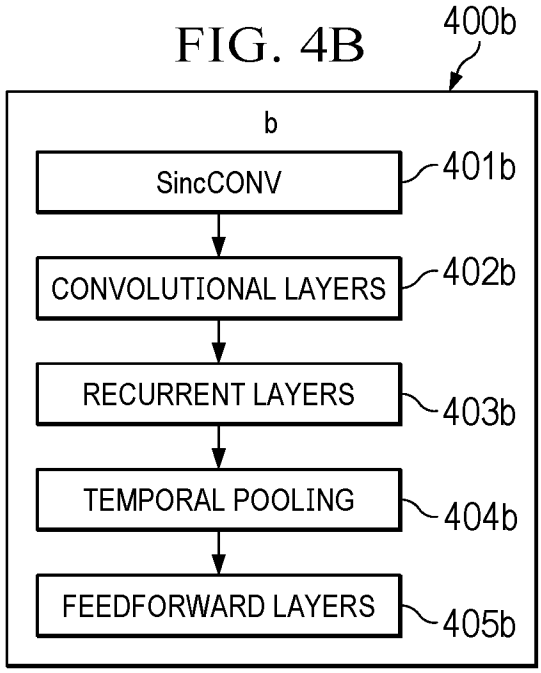


FIG. 4B



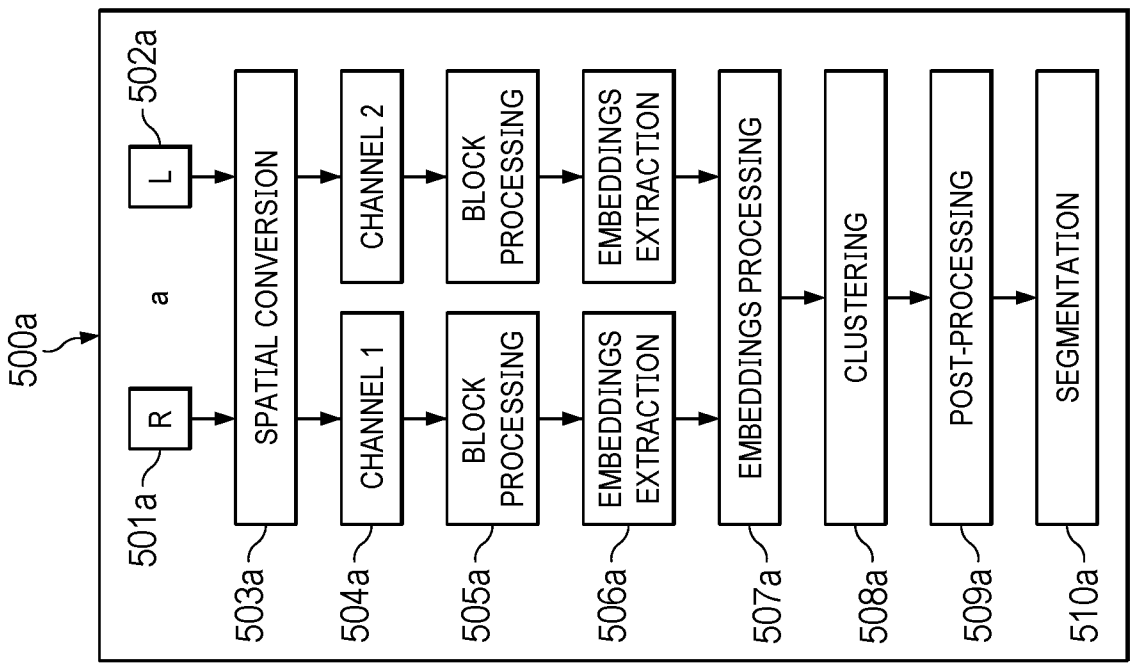

500a
501a
R
a
L
502a
SPATIAL CONVERSION
503a
CHANNEL 1
CHANNEL 2
504a
BLOCK PROCESSING
BLOCK PROCESSING
505a
EMBEDDINGS EXTRACTION
EMBEDDINGS EXTRACTION
506a
EMBEDDINGS PROCESSING
507a
CLUSTERING
508a
POST-PROCESSING
509a
SEGMENTATION
510a

FIG. 5A

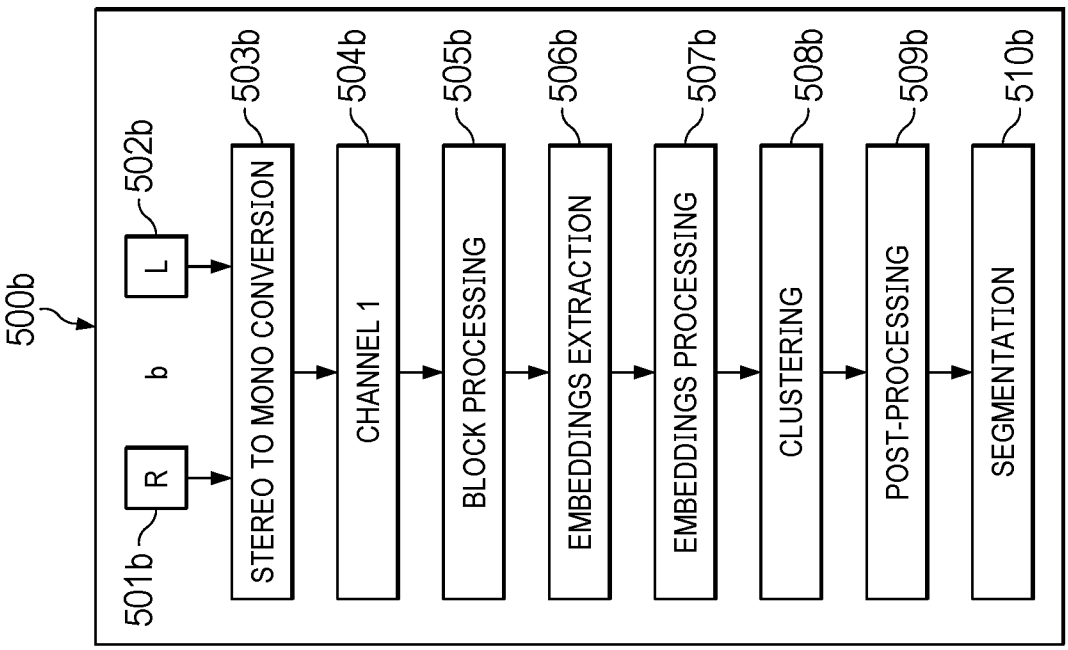

500b
501b
R
b
L
502b
STEREO TO MONO CONVERSION
503b
CHANNEL 1
504b
BLOCK PROCESSING
505b
EMBEDDINGS EXTRACTION
506b
EMBEDDINGS PROCESSING
507b
CLUSTERING
508b
POST-PROCESSING
509b
SEGMENTATION
510b

FIG. 5B

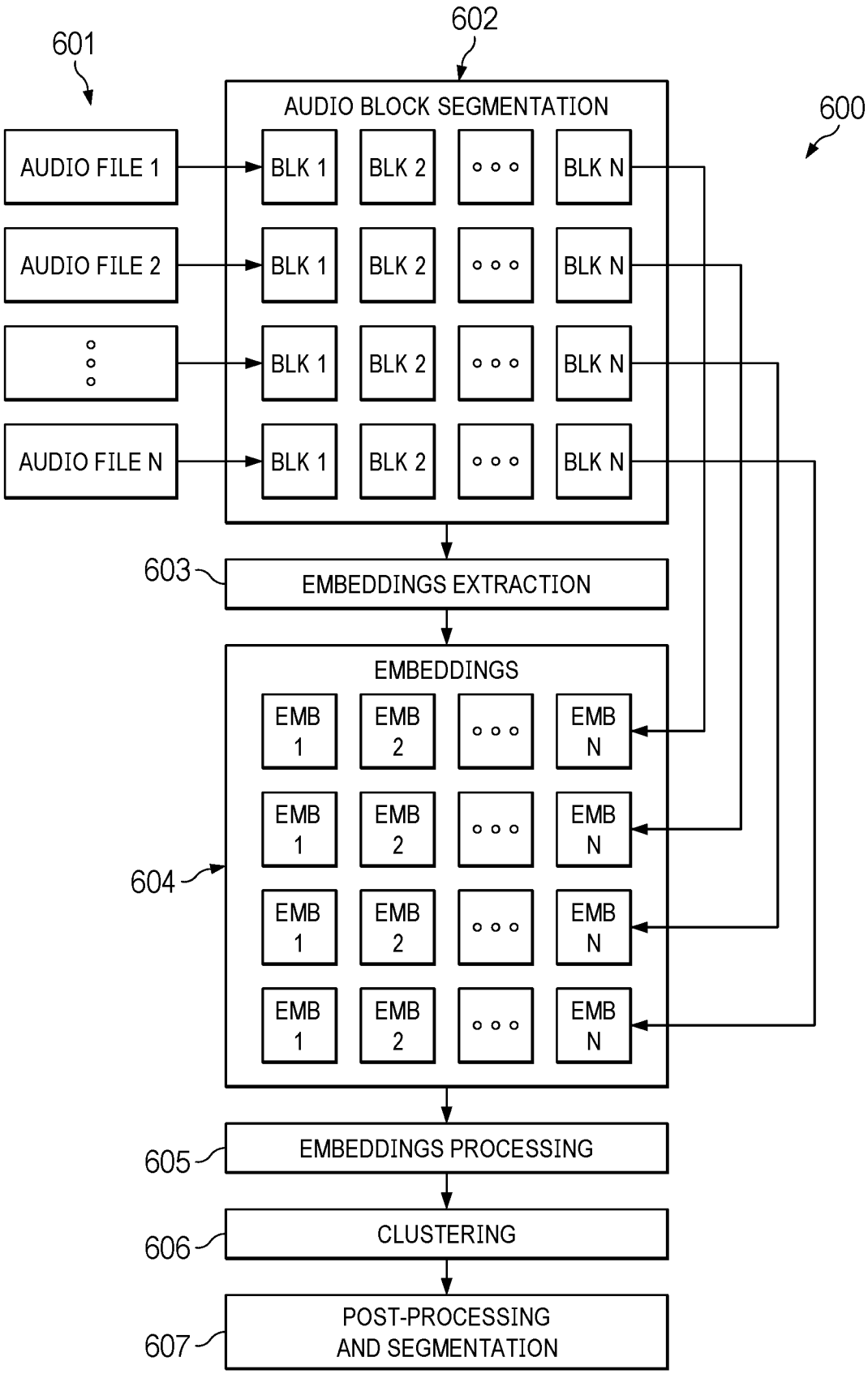

601
602
600
AUDIO BLOCK SEGMENTATION
AUDIO FILE 1
AUDIO FILE 2
AUDIO FILE N
BLK 1
BLK 2
BLK N
603
EMBEDDINGS EXTRACTION
EMBEDDINGS
604
EMB 1
EMB 2
EMB N
605
EMBEDDINGS PROCESSING
606
CLUSTERING
607
POST-PROCESSING AND SEGMENTATION

FIG. 6

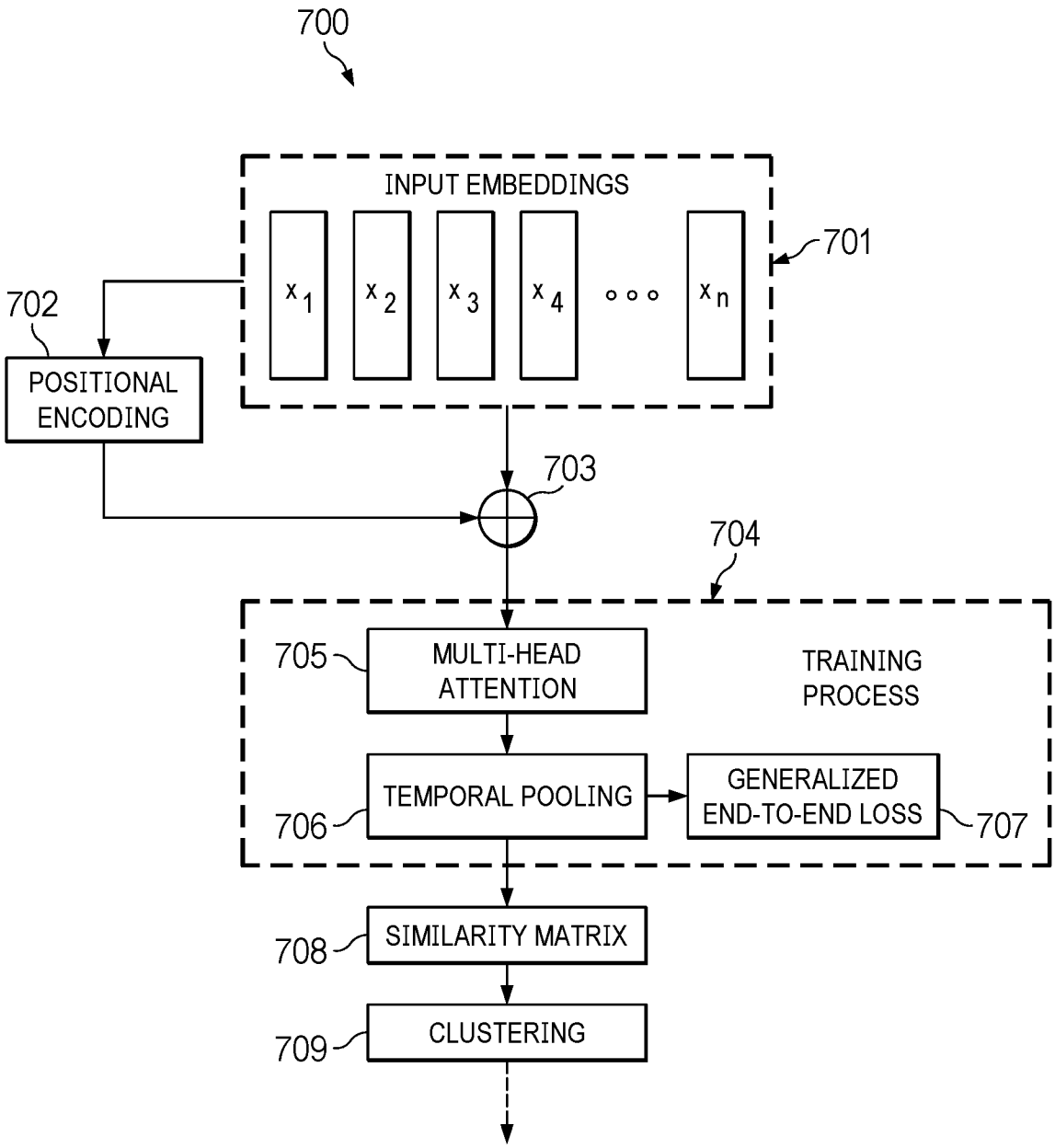

700
INPUT EMBEDDINGS
x 1
x 2
x 3
x 4
o o o
x n
701
702
POSITIONAL ENCODING
703
704
705
MULTI-HEAD ATTENTION
TRAINING PROCESS
TEMPORAL POOLING
GENERALIZED END-TO-END LOSS
706
707
708
SIMILARITY MATRIX
709
CLUSTERING

FIG. 7

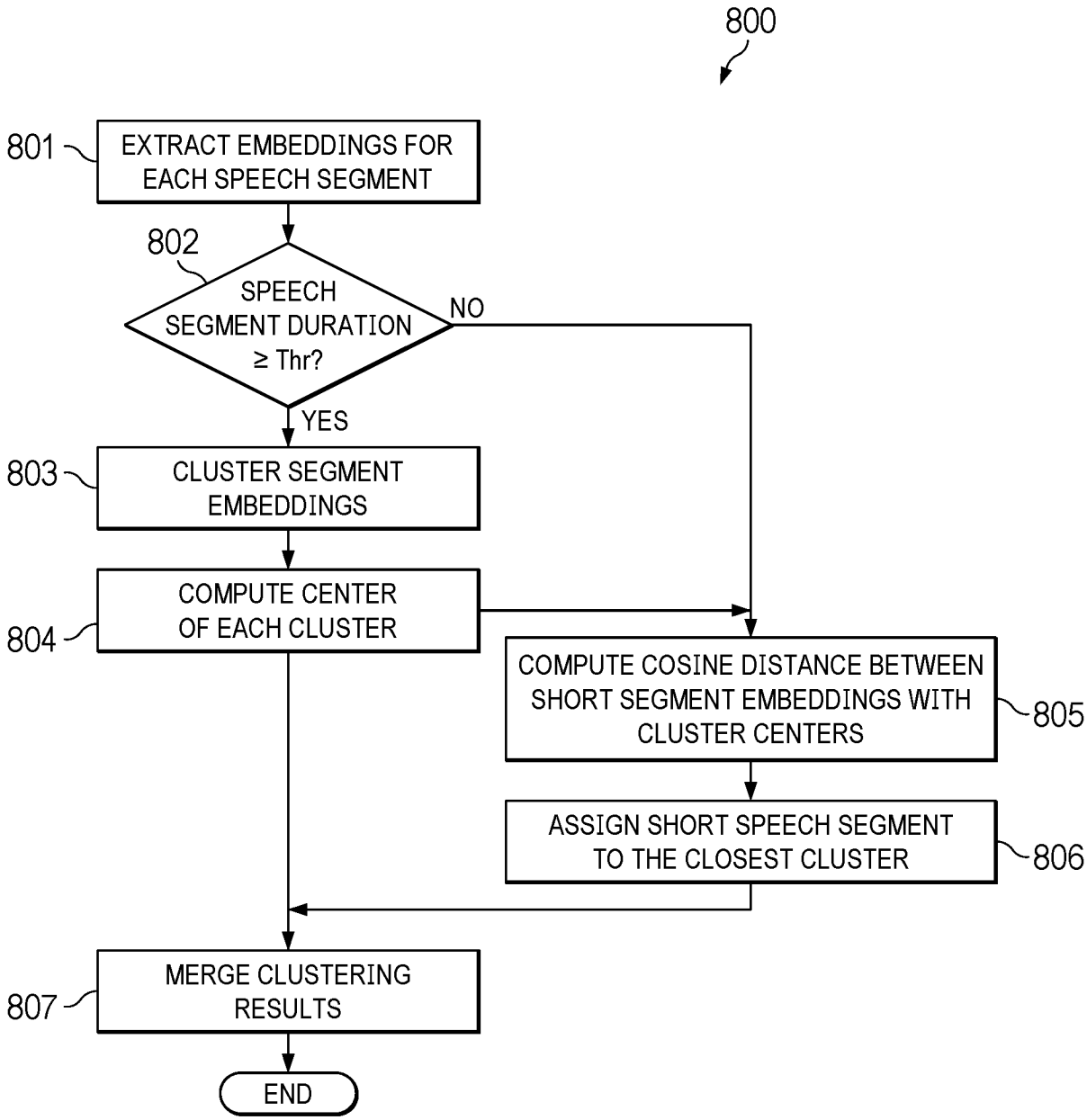

800
801
EXTRACT EMBEDDINGS FOR EACH SPEECH SEGMENT
802
SPEECH SEGMENT DURATION ≥ Thr?
NO
YES
803
CLUSTER SEGMENT EMBEDDINGS
804
COMPUTE CENTER OF EACH CLUSTER
805
COMPUTE COSINE DISTANCE BETWEEN SHORT SEGMENT EMBEDDINGS WITH CLUSTER CENTERS
806
ASSIGN SHORT SPEECH SEGMENT TO THE CLOSEST CLUSTER
807
MERGE CLUSTERING RESULTS
END

FIG. 8

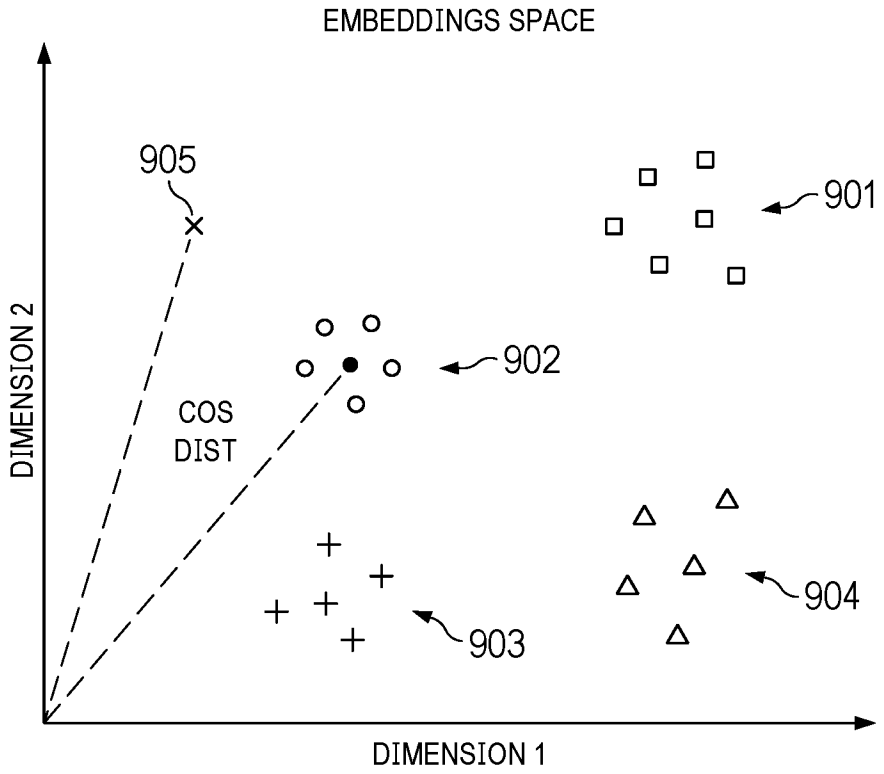

EMBEDDINGS SPACE
905
901
902
COS
DIST
DIMENSION 2
904
903
DIMENSION 1

FIG. 9

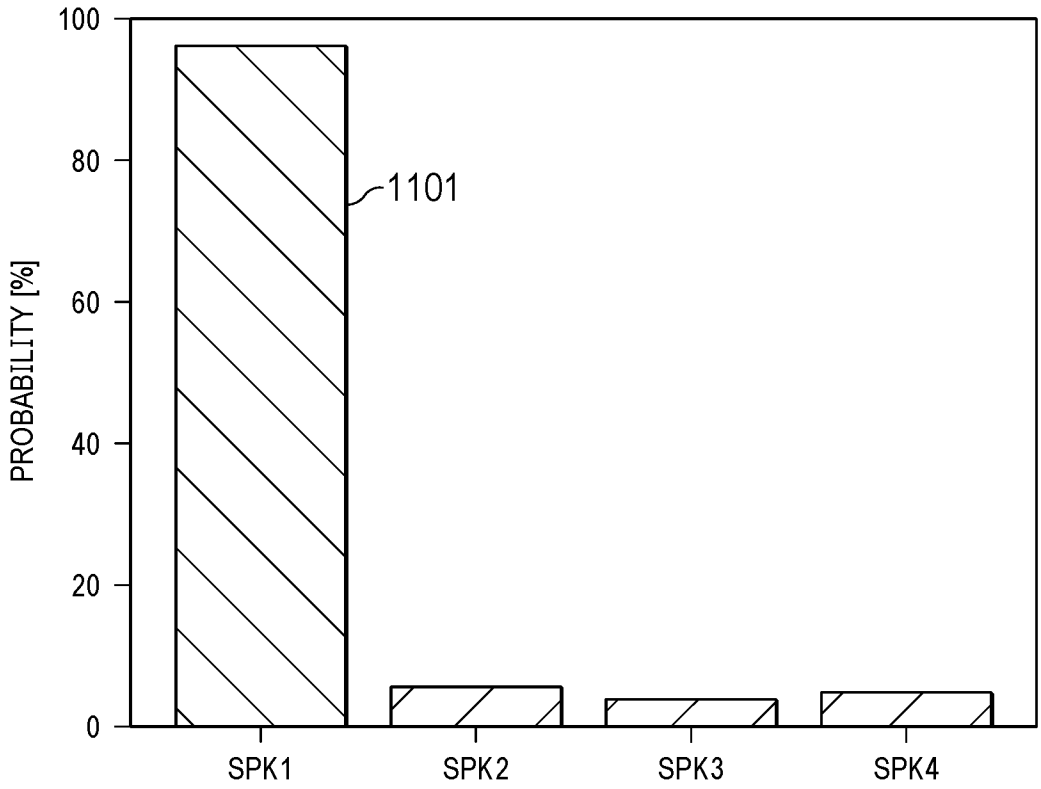

100
80
60
40
20
0
PROBABILITY [%]
1101
SPK1
SPK2
SPK3
SPK4

FIG. 11

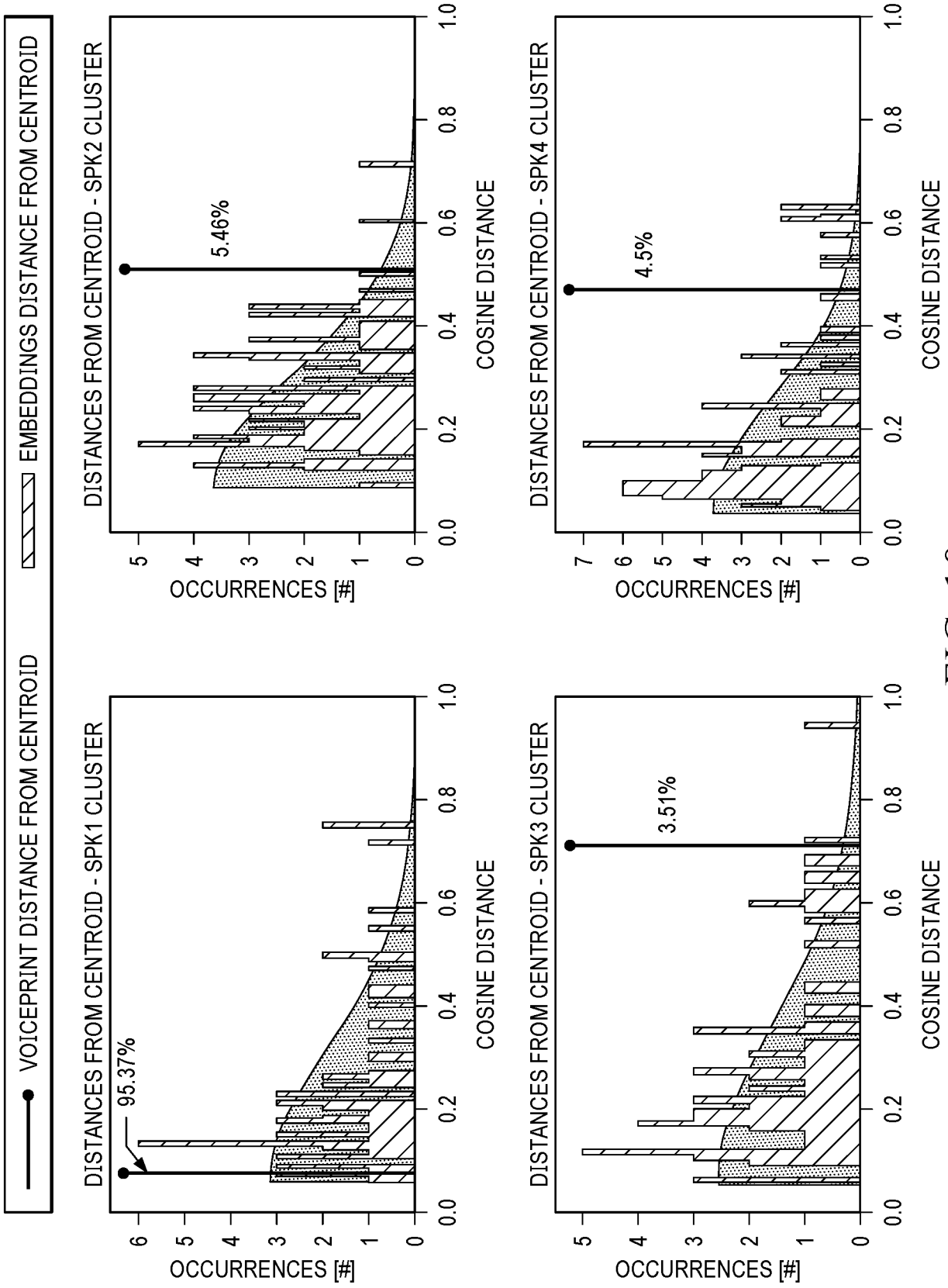

VOICEPRINT DISTANCE FROM CENTROID
EMBEDDINGS DISTANCE FROM CENTROID
DISTANCES FROM CENTROID - SPK1 CLUSTER
95.37%
DISTANCES FROM CENTROID - SPK2 CLUSTER
5.46%
DISTANCES FROM CENTROID - SPK3 CLUSTER
3.51%
DISTANCES FROM CENTROID - SPK4 CLUSTER
4.5%
OCCURRENCES [#]
COSINE DISTANCE

FIG. 10

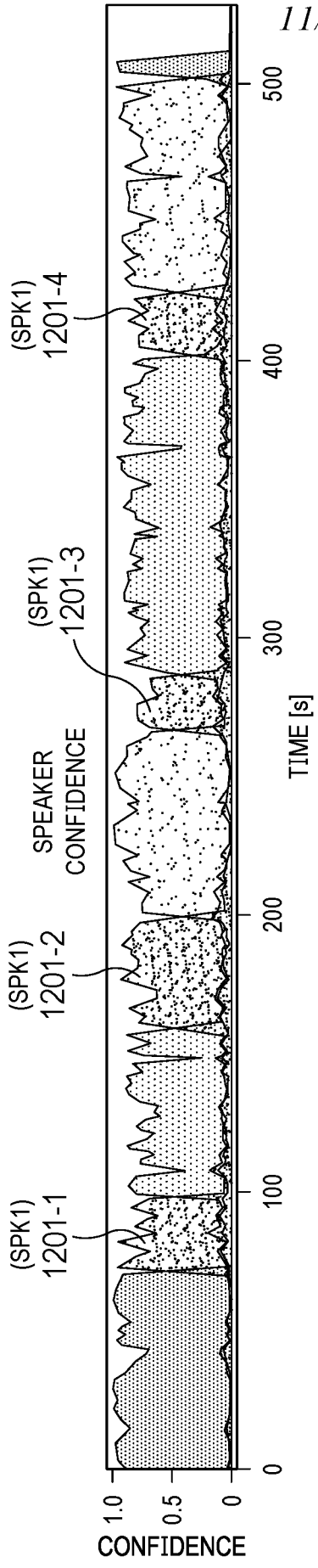

(SPK1)
1201-1
(SPK1)
1201-2
(SPK1)
1201-3
(SPK1)
1201-4
SPEAKER
CONFIDENCE
CONFIDENCE
1.0
0.5
0
0
100
200
300
400
500
TIME [s]

FIG. 12

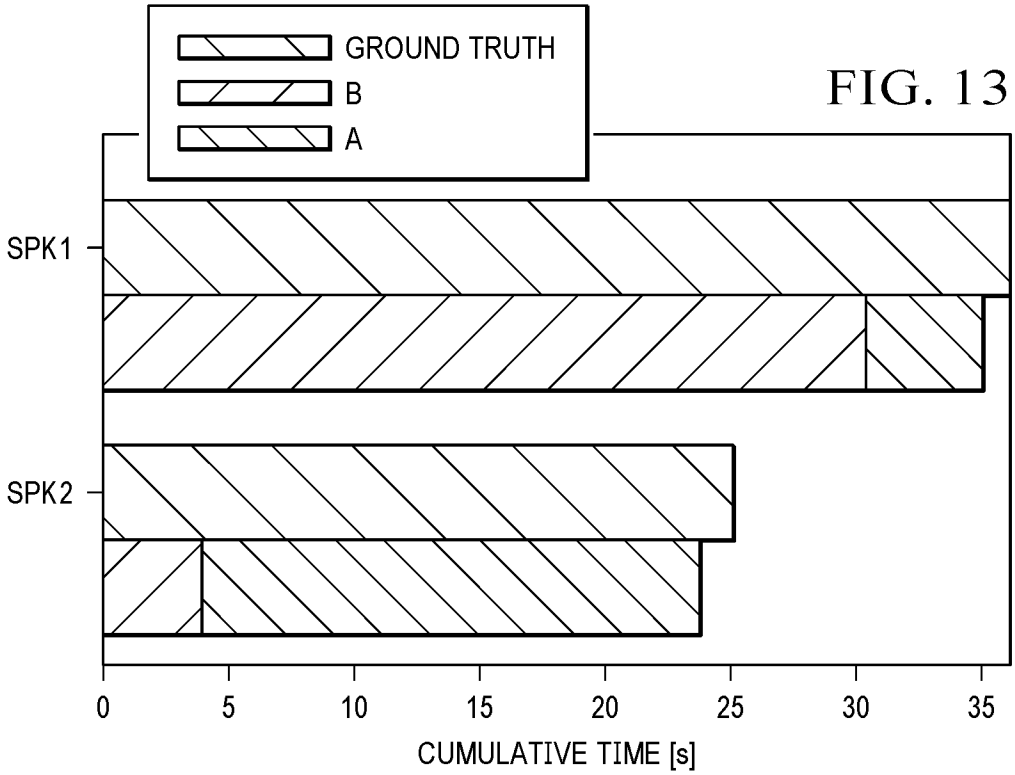

GROUND TRUTH
B
A
SPK1
SPK2
0
5
10
15
20
25
30
35
CUMULATIVE TIME [s]

FIG. 13

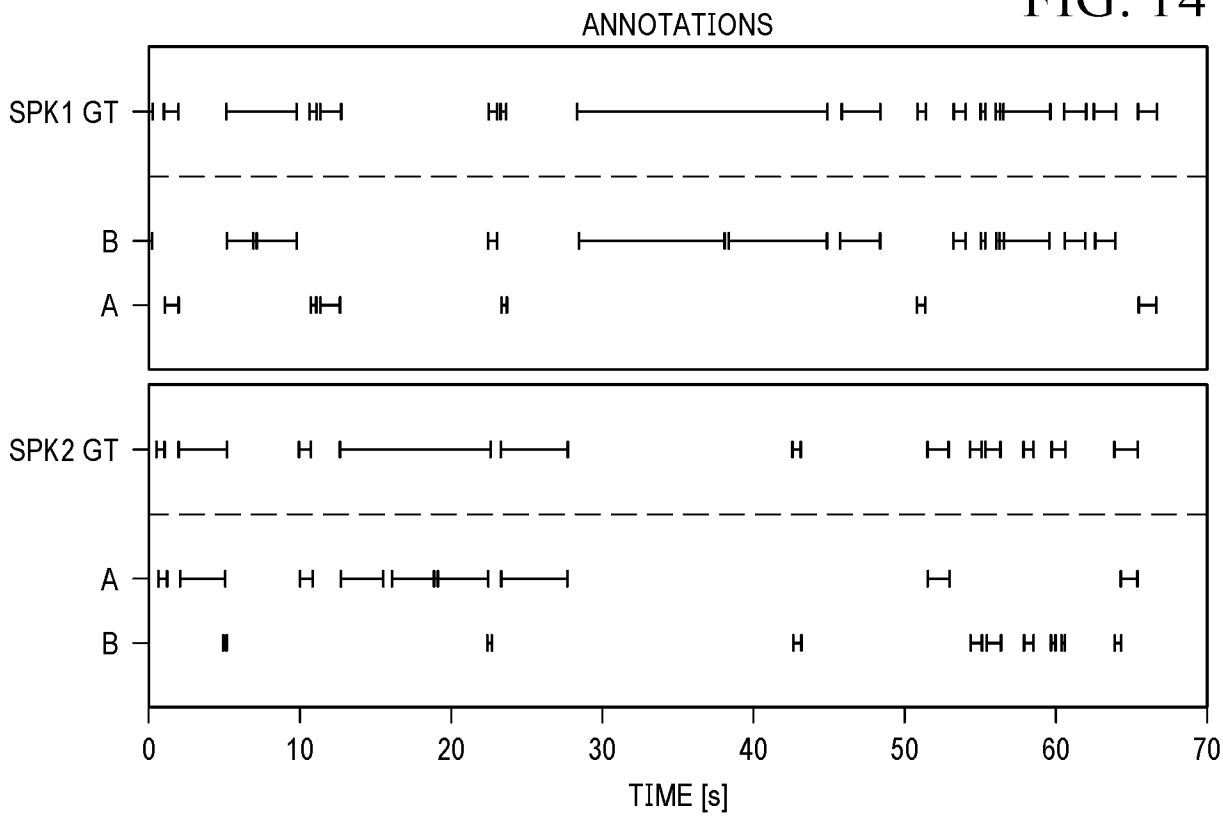

ANNOTATIONS
SPK1 GT
B
A
SPK2 GT
A
B
0
10
20
30
40
50
60
70
TIME [s]

FIG. 14

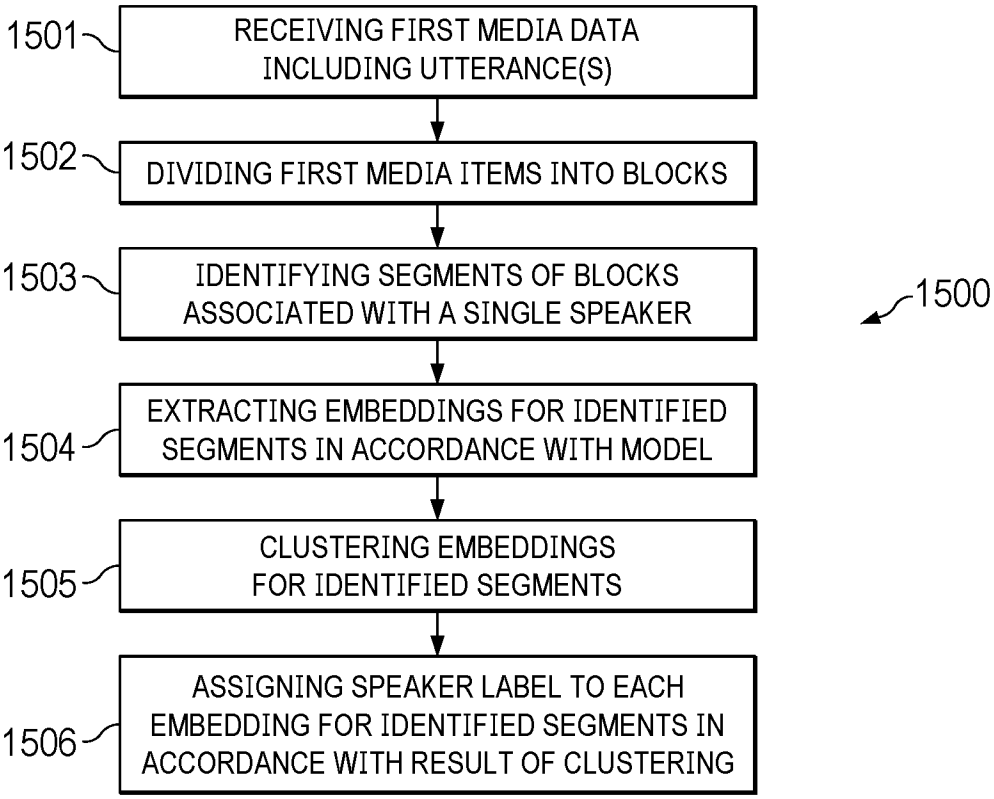

1501
RECEIVING FIRST MEDIA DATA INCLUDING UTTERANCE(S)
1502
DIVIDING FIRST MEDIA ITEMS INTO BLOCKS
1503
IDENTIFYING SEGMENTS OF BLOCKS ASSOCIATED WITH A SINGLE SPEAKER
1500
1504
EXTRACTING EMBEDDINGS FOR IDENTIFIED SEGMENTS IN ACCORDANCE WITH MODEL
1505
CLUSTERING EMBEDDINGS FOR IDENTIFIED SEGMENTS
1506
ASSIGNING SPEAKER LABEL TO EACH EMBEDDING FOR IDENTIFIED SEGMENTS IN ACCORDANCE WITH RESULT OF CLUSTERING

FIG. 15

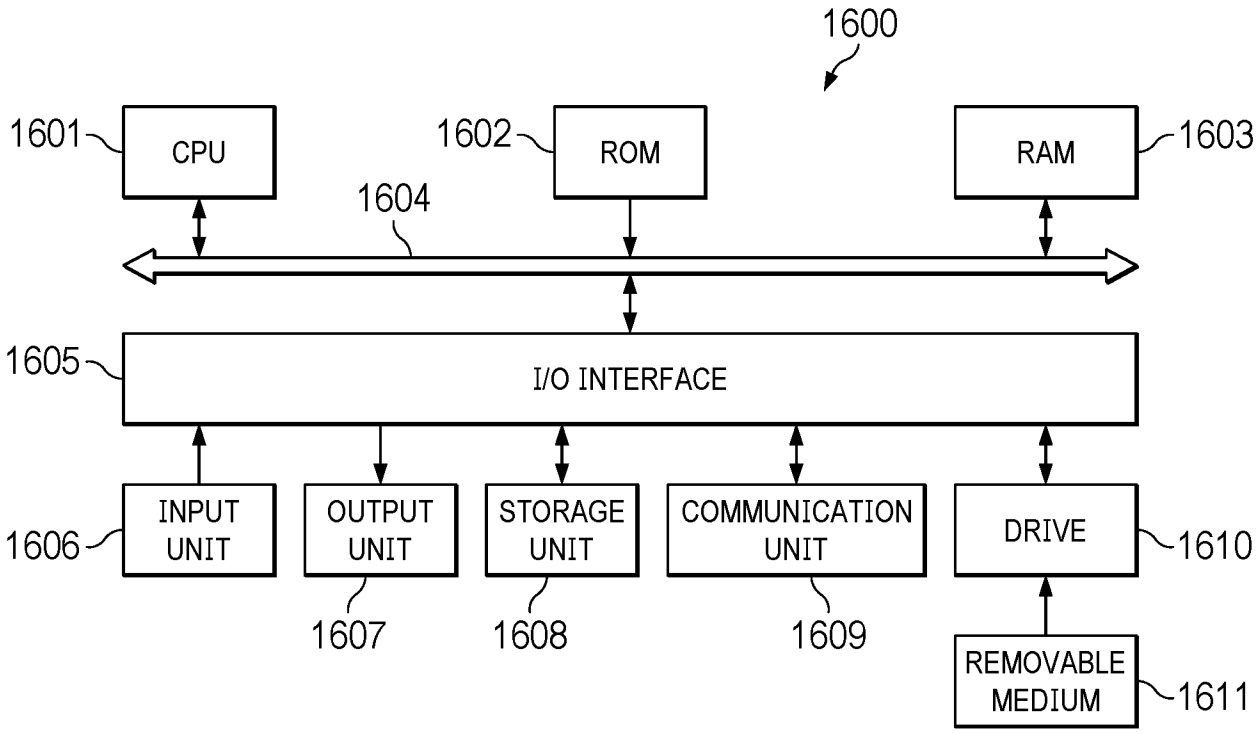

1600
1601
CPU
1602
ROM
1603
RAM
1604
1605
I/O INTERFACE
1606
INPUT UNIT
OUTPUT UNIT
STORAGE UNIT
COMMUNICATION UNIT
DRIVE
1610
1607
1608
1609
REMOVABLE MEDIUM
1611

FIG. 16

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/026552

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L25/84 G10L17/02 G10L17/06 G10L17/18
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2016/217792 A1 (GORODETSKI ALEX [IL] ET AL) 28 July 2016 (2016-07-28)	1,2,4-8, 10,11, 13-23
Y	abstract; figures	3,12
A	paragraphs [0022] - [0065]	9
Y	US 2020/349230 A1 (YOSHIOKA TAKUYA [US] ET AL) 5 November 2020 (2020-11-05)	3,12
A	abstract; figures 1-13, 17-19	1,2,
	paragraphs [0092] - [0118]	4-11,
	paragraphs [0131] - [0143]	13-23
A	WO 2020/068056 A1 (GOOGLE LLC [US]) 2 April 2020 (2020-04-02)	1-23
	abstract; figures 1-13	
	paragraphs [0001] - [0028]	
	paragraphs [0071] - [0114]	

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
8 July 2022	19/07/2022

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Képesi, Marián

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2022/026552

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2016217792 A1	28-07-2016	US 2016217792 A1	28-07-2016
		US 2016217793 A1	28-07-2016
		US 2018211670 A1	26-07-2018
		US 2018218738 A1	02-08-2018
		US 2020349956 A1	05-11-2020
US 2020349230 A1	05-11-2020	AU 2020265509 A1	07-10-2021
		CA 3132837 A1	05-11-2020
		CN 113874936 A	31-12-2021
		EP 3963574 A1	09-03-2022
		KR 20220002326 A	06-01-2022
		SG 11202109373T A	29-09-2021
		US 2020349230 A1	05-11-2020
		WO 2020222925 A1	05-11-2020
WO 2020068056 A1	02-04-2020	CN 112262431 A	22-01-2021
		EP 3655947 A1	27-05-2020
		EP 4009321 A1	08-06-2022
		JP 2021533397 A	02-12-2021
		KR 20210008520 A	22-01-2021
		US 2020342857 A1	29-10-2020
		US 2021217411 A1	15-07-2021
		WO 2020068056 A1	02-04-2020